PRIOR ART INFORMATION LIST

Your case No.	
Our case No.	20574

Inventor, Patent number, Country, Author, Title, Number of Document	Issue date	Concise Explanation of the Relevance (indication of page, column, line, figure of the relevant portion)
JP-A-8-153121	June 11, 1996	Abstract
Atsuo KAWAI "AN AUTOMATIC DOCUMENT CLASSIFICATION METHOD BASED ON A SEMANTIC CATEGORY FREQUENCY ANALYSIS", Transactions of Information Processing Society of Japan (Information Processing Society of Japan) vol.33, No.9	Sep. 15, 1992	English translation of p.1114 Summary, 1(Instruction), and p.1119 3.3(Evaluation criteria) ~ p.1121 3.4(Evaluation of precision)/ Fig.5-1,5-2,and 6

(19) JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **08153121 A**

(43) Date of publication of application: **11.06.96**

(51) Int. Cl **G06F 17/30**
G06F 12/00

(21) Application number: **07231033**

(22) Date of filing: **08.09.95**

(30) Priority: **30.09.94 JP 06236444**

(71) Applicant: **HITACHI LTD**

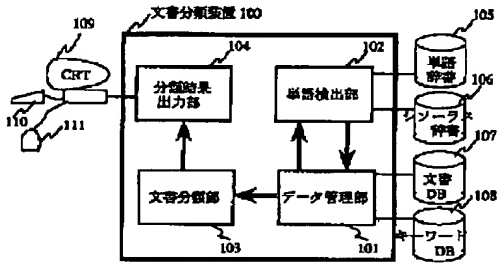
(72) Inventor: **MORITA TAKAKO**
TONO JUNICHI
MATSUDA YOSHIKI
HASHIMOTO TETSUYA

(54) **METHOD AND DEVICE FOR DOCUMENT INFORMATION CLASSIFICATION**

(57) Abstract:

PURPOSE: To provide a method and device for document information classification which classify a document group without depending upon a prescribed classification system by using a key word given to the document group or a word appearing in a document and rearranges classification results hierarchically.

CONSTITUTION: A data management part 101 manages the document group in a document DB 107 and a group of key words, given to respective documents, in a key word DB 108. A document classification part 103 classifies the documents on the basis of the individual key words and stores them in folders. Then, folders having similar document groups are integrated. For the integration, it is judged whether the integration is effective or not. It is judged whether or not further classifications can be made in folders that are left without being integrated, thereby generating a hierarchical classification system. The classification results are outputted on a CRT 109 by a classification output part 104 to provide an environment wherein a user can read the classification results out.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平8－153121

(43)公開日　平成８年(1996)６月11日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ			技術表示箇所
Ｇ０６Ｆ	17/30						
	12/00	５２０　Ｅ	7623－5B				
			9194－5L	Ｇ０６Ｆ	15/ 403	３２０　Ｃ	
			9194－5L		15/ 401	３１０　Ｄ	

審査請求　未請求　請求項の数25　ＯＬ　（全 37 頁）

(21)出願番号　特願平7－231033

(22)出願日　平成７年(1995)９月８日

(31)優先権主張番号　特願平6－236444

(32)優先日　平６(1994)９月30日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地

(72)発明者　森田　貴子
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　東野　純一
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　松田　芳樹
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　小川　勝男

最終頁に続く

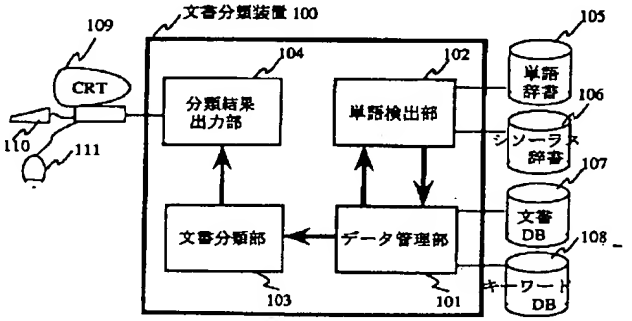
(54)【発明の名称】　文書情報分類方法および文書情報分類装置

(57)【要約】

【目的】　本発明の目的は、文書群に付与されたキーワードおよび文書中の出現単語を手がかりとして、規定の分類体系に依存せずに文書群を分類し、分類結果を階層的に整理する文書情報分類方法および文書情報分類装置を提供することにある。

【構成】　文書ＤＢ１０７の文書群とキーワードＤＢ１０８の各文書に付与されたキーワード群をデータ管理部１０１で管理する。文書分類部１０３では、個々のキーワードに基づいて文書をフォルダに分類して格納する。次に、類似した文書群を持つフォルダを統合する。統合する際には統合が有効かどうかの判断を行う。統合されたフォルダ内や、統合されずに残ったフォルダ内は、細分類が可能かどうかを判断し、階層的な分類体系を作り出す。分類結果は分類結果出力部１０４でＣＲＴ１０９に出力し、ユーザが分類結果を読み出すことができる環境を提供する。

図１



1

【特許請求の範囲】

【請求項１】予め蓄積された複数の文書情報のそれぞれを所定の分類に分類する文書情報分類方法において、

上記文書情報および上記文書情報の各文書情報におけるキーワード群を蓄積しておき、

蓄積された各キーワード毎に、キーワードを含む上記文書情報を単一キーワードフォルダにまとめ、

上記単一キーワードフォルダ内の文書同士を比較することにより、上記単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成し、

予め蓄積された上記複数の文書情報を上記関連キーワードフォルダに分類することを特徴とする文書情報分類方法。

【請求項２】請求項１に記載の文書情報分類方法において、

上記関連キーワードフォルダ同士を結合し、

上記複数の文書情報を新たな関連キーワードフォルダに分類することを特徴とする文書情報分類方法。

【請求項３】請求項２に記載の文書情報分類方法において、

予め関連キーワードフォルダの数を指定しておき、

上記関連キーワードフォルダ同士を結合する際、指定された関連キーワードフォルダの数になるまで結合を繰り返すことを特徴とする文書情報分類方法。

【請求項４】請求項２に記載の文書情報分類方法において、

上記新たな関連キーワードフォルダ内の各文書の内容により、上記新たな関連キーワードフォルダを存続させるか否か確認することを特徴とする文書情報分類方法。

【請求項５】請求項１に記載の文書情報分類方法において、

上記関連キーワードフォルダ内の文書情報同士を比較することにより、上記文書情報を第２の単一キーワードフォルダにまとめ、

上記第２の単一キーワードフォルダ内の文書情報同士を比較することにより、上記第２の単一キーワードフォルダ同士を結合して、第２の関連キーワードフォルダを作成し、

上記関連キーワードフォルダ内の文書情報を上記第２の関連キーワードフォルダに分類することを特徴とする文書情報分類方法。

【請求項６】請求項５に記載の文書情報分類方法において、

予め上記第２の関連キーワードフォルダ内に分類される文書情報の数を定めておき、

上記第２の関連キーワードフォルダに分類するステップは、予め定められた数になるまで上記文書情報を分類することを特徴とする文書情報分類方法。

【請求項７】請求項１に記載の文書情報分類方法において、

2

予め蓄積された上記複数の文書情報中に現れる言葉を抽出し、

抽出された上記言葉を上記キーワードとすることを特徴とする文書情報分類方法。

【請求項８】請求項１に記載の文書情報分類方法において、

上記単一キーワードフォルダ内の文書同士を比較して、一致する文書情報の数が所定以上の単一キーワードフォルダ同士を結合して上記関連キーワードフォルダを作成することを特徴とする文書情報分類方法。

【請求項９】請求項１に記載の文書情報分類方法において、

上記キーワードの上記文書情報中での出現頻度および出現位置のうちいずれか一方を用いることにより上記関連キーワードフォルダを作成することを特徴とする文書情報分類方法。

【請求項１０】請求項１に記載の文書情報分類方法において、

上記関連キーワードフォルダ内の各文書情報の特徴ベクトルを算出して、算出された各特徴ベクトルの平均ベクトルを求め、

求められた上記平均ベクトルと特徴ベクトルとの差が所定以上の文書情報を上記関連キーワードフォルダ内で再分割することを特徴とする文書情報分類方法。

【請求項１１】請求項１に記載の文書情報分類方法において、

上記関連キーワードフォルダ内の各文書情報の特徴ベクトルを算出し、

算出した上記特徴ベクトルを用いて上記関連キーワードフォルダの結合の可否を判定することを特徴とする文書情報分類方法。

【請求項１２】予め複数の文書情報および上記文書情報の各文書情報におけるキーワード群を蓄積しておく記憶手段を有し、上記文書情報を分類する文書情報分類装置において、

上記記憶手段に蓄積された各キーワード毎に、キーワードを含む上記文書情報を単一キーワードフォルダにまとめる単一キーワード処理手段と、

上記単一キーワードフォルダ内の文書情報同士を比較することにより、上記単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成する関連キーワード処理手段とを有し、

予め蓄積された上記複数の文書情報を上記関連キーワードフォルダに分類することを特徴とする文書情報分類装置。

【請求項１３】請求項１２に記載の文書情報分類装置において、

上記関連キーワードフォルダ同士を結合する関連キーワード結合手段を有し、上記複数の文書情報を新たな関連キーワードフォルダに分類することを特徴とする文書情

報分類装置。

【請求項１４】請求項１３に記載の文書情報分類装置において、

上記新たな関連キーワードフォルダ内の各文書情報の内容により、上記新たな関連キーワードフォルダを存続させるか否か確認する確認手段を有することを特徴とする文書情報分類装置。

【請求項１５】請求項１２に記載の文書情報分類装置において、

上記関連キーワードフォルダ内の文書情報同士を比較することにより、上記文書情報を第２の単一キーワードフォルダにまとめる第２の単一キーワード処理手段と、

上記第２の単一キーワードフォルダ内の文書情報同士を比較することにより、上記第２の単一キーワードフォルダ同士を結合して、第２の関連キーワードフォルダを作成する第２の関連キーワード処理手段とを有し、

上記関連キーワードフォルダ内の文書情報を上記第２の関連キーワードフォルダに分類することを特徴とする文書情報分類装置。

【請求項１６】請求項１２に記載の文書情報分類装置において、

予め蓄積された上記複数の文書情報中に現れる言葉を抽出する抽出手段と、

抽出された上記言葉を上記キーワードとするキーワード作成手段を有することを特徴とする文書情報分類装置。

【請求項１７】請求項１２に記載の文書情報分類装置において、

上記関連キーワードフォルダ処理手段は、上記単一キーワードフォルダ内の文書情報同士を比較して、一致する文書情報の数が所定以上の単一キーワードフォルダ同士を結合して上記関連キーワードフォルダを作成することを特徴とする文書情報分類装置。

【請求項１８】請求項１２に記載の文書情報分類装置において、

上記関連キーワードフォルダ処理手段は、上記キーワードの上記文書情報中での出現頻度および出現位置のうちいずれか一方を用いることにより上記関連キーワードフォルダを作成することを特徴とする文書情報分類装置。

【請求項１９】請求項１２に記載の文書情報分類装置において、

上記関連キーワードフォルダ内の各文書の特徴ベクトルを算出して、算出された各特徴ベクトルの平均ベクトルを求める平均ベクトル算出手段と、

求められた上記平均ベクトルと特徴ベクトルとの差が所定以上の文書情報を上記関連キーワードフォルダ内で再分割する再分割手段とを有することを特徴とする文書情報分類装置。

【請求項２０】請求項１２に記載の文書情報分類装置において、

上記関連キーワードフォルダ内の各文書の特徴ベクトル

を算出する特徴ベクトル算出手段と、

算出した上記特徴ベクトルを用いて上記関連キーワードフォルダの結合の可否を判定する結合判定手段とを有することを特徴とする文書情報分類装置。

【請求項２１】予め複数の文書情報および上記文書情報の各文書情報におけるキーワード群を蓄積しておく記憶手段を有し、上記文書情報を分類する文書情報分類装置において、

上記記憶手段に蓄積された各キーワード毎に、キーワードを含む上記文書情報を単一キーワードフォルダにまとめる単一キーワード処理手段と、

上記単一キーワードフォルダ内の文書情報同士を比較することにより、上記単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成する関連キーワード処理手段と、

操作者が指定した上記関連キーワードフォルダ同士を結合する手段を有し、

予め蓄積された上記複数の文書情報を上記関連キーワードフォルダに分類することを特徴とする文書情報分類装置。

【請求項２２】請求項１２に記載の文書情報分類装置において、

上記単一キーワード処理手段は、操作者が選択した文字列を含む文書情報を単一キーワードフォルダにまとめることを特徴とする文書情報分類装置。

【請求項２３】請求項２２に記載の文書情報分類装置において、

文字列の階層構造を示すシソーラス辞書を有し、

上記シソーラス辞書を用いて、上記関連キーワードフォルダの階層関係を構築する階層関係構築手段とを有することを特徴とする文書情報分類装置。

【請求項２４】請求項２２に記載の文書情報分類装置において、

上記関連キーワードフォルダに含まれる文書情報に基づいて関連キーワードフォルダ同士の類似度を定める手段と、

上記類似度に応じて、操作者が選択した関連キーワードフォルダに類似する関連キーワードフォルダを選択する手段を有することを特徴とする文書情報分類装置。

【請求項２５】請求項２４に記載の文書情報分類装置において、

上記類似度に応じて、操作者が選択した文書情報と同じ関連キーワードフォルダに属し、類似する文書情報を上記関連キーワードフォルダから取り除く手段とを有することを特徴とする文書情報分類装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、計算機上に蓄積された文書ファイルを階層的に分類する文書情報分類方法および文書情報分類装置に関する。特に、情報分類システム

および文書ファイリングシステム等に適用されるものである。

【０００２】

【従来の技術】計算機ネットワークの整備によって、情報検索サービスや電子メールを利用して、情報収集が行える環境が整いつつある。計算機ネットワークを介して、新しい情報が次々に到着し、話題は時々刻々変化する。最新情報の有用性は十分認識されているが、次の問題によって有効活用されていないのが現状である。

【０００３】（１）所望の情報を選択するための検索式を論理式で入力しなければならない。このことは一般ユーザには困難である。

【０００４】（２）所望する情報を得るために適切な検索式を作成することが困難である。

【０００５】（３）収集した情報数が多くなると読み切れず、収集した中で必要な情報だけを選択して読むことができない。

【０００６】「２１世紀の情報化社会」（日経バイト、１９９１年１１月、３２０～３３１ページ）に記載されている広域情報サーバは、（１）、（２）の問題点を関連性フィードバックにより解決している。関連性フィードバックとは、ユーザが提示した検索条件に基づいて検索を行い、検索結果の中からユーザが所望の情報を選択すると、選択した情報を検索条件にフィードバックして、検索条件を改善するものである。はじめにユーザが提示した検索条件が不適切であっても、後に学習が行われて検索条件が洗練されるという手法である。

【０００７】「情報のブロードキャッチシステム」（情報処理学会 情報メディア研究会１３-６、グループウェア研究会４-６報告、１９９３年１０月２８日、３７ページ～４４ページ）に記載されているシステムでは、ユーザが自分の興味をキーワードとしてシステムに登録しておくと、これに適合する文書を収集する。論理式の入力を必要とせず、キーワードの登録だけでよいことから、（１）の問題点が解決できる。さらに、収集した文書数が一定数を超えると、文書群を自動分類している。文書の内容をキーワードの出現頻度の並びによるワードベクトルで表現し、ワードベクトル間の類似度を計算して類似する文書をまとめ、文書を分類する。

【０００８】また、特開平５-２８１９８号公報記載の文書情報検索装置は、各文書に付与された分類や文書中に出現する単語といった検索情報を手がかりとして文書検索を行う。ユーザが検索したい分野や単語などの検索データを入力すると、内容を推測し自動的に検索式を作成して検索を行うことにより、（１）の問題を解決している。検索式の作成に際しては、あらかじめ登録してある検索論理式のうち、最適な論理式を選択する。さらに、検索された文書群は検索データとの類似度を算出し、その値を使って整列することで、（３）の問題に対処している。

【０００９】また、これらの従来例では、１つの視点でしか文書分類できない。そのため、検索漏れを起こすとの問題もある。

【００１０】

【発明が解決しようとする課題】前述の「２１世紀の情報化社会」に記載された広域情報サーバでは、関連性フィードバックという方法によって、ユーザが指定した検索条件だけではなく、ユーザの希望にかなった実情報を利用して、システムが検索条件を改善し、（１）、（２）の問題を解決した。

【００１１】「情報のブロードキャッチシステム」では、検索式を入力するのではなく、キーワードを登録するという方法を採用している。しかし、ユーザが自分の興味を的確に反映するキーワードを登録するのは困難なため、（２）の問題は解決し切れていない。また、文書群の自動分類処理は逐次的に起動されることが前提なので、文書の到着順序に依存した分類になる。情報の内容は時々刻々変化するので、過去のある時点における分類体系に依存し続けると、有効な分類結果は生成できず、（３）の解決は困難になる。

【００１２】上述したように、文書の収集、検索処理ではユーザが自分の興味を的確に反映したキーワードを設定する作業が困難であるという問題がある。

【００１３】この検索処理に対する問題と比較して、文書の分類処理でも分類の手がかりとなる分類体系をあらかじめ設定する作業においても同様の問題が生じる。つまり、一つの分類体系に依存し続けると、内容の変化に対応することができず、不適切な分類結果になり得るという問題が生じる。

【００１４】一方、特開平５-２８１９８号公報記載の文書情報検索装置では、検索結果を有効なものから順に整列して検索式に類似する文書から見られるように工夫している。しかし、何番目までが有効な情報かを判断するには、ユーザは文書の内容を確認せざるを得ないため、（３）の問題を解決できるとはいえない。

【００１５】つまり、検索結果を一階層に整列しだけでは、類似した内容の文書をまとめて見たり、まとめて読み飛ばしたりすることができないという問題がある。

【００１６】これらの問題を解決するものとして、特開平５－３２４７２６号公報がある。この従来例では、あるキーワードに着目して文書内にそのキーワードが存在するか否かにより文書を分類する。分類された文書に同様の処理を施すことにより階層的に分類していく。

【００１７】しかし、この従来例でも複数分野に関連する文書を１つの視点でしか分類できないとの問題点を有する。

【００１８】また、本従来例を新聞などの一般の文書データに適用すると階層が深くなり、分類数が爆発的に増加するとの問題もある。

【００１９】本発明の目的は上述した問題を解決するた

めに、既定の分類体系にとらわれることなく文書を自動的に分類し、分類結果を階層的に整理する文書分類方法および文書分類装置を提供することにある。

【００２０】

【課題を解決するための手段】上記の目的を達成するために本発明は、予め複数の文書および文書の各文書におけるキーワード群を蓄積しておき、蓄積された各キーワード毎に、キーワードを含む文書を単一キーワードフォルダにまとめ、単一キーワードフォルダ内の文書同士を比較することにより、単一キーワードフォルダ同士を結合して関連キーワードフォルダを作成し、予め蓄積された複数の文書を関連キーワードフォルダに分類することを特徴とする。

【００２１】本発明の文書情報分類方法および文書情報分類装置では、蓄積された複数の文書と、各文書に人手で付与したキーワードと各文書中から自動抽出した単語をまとめたキーワード群と、キーワード群中の個々のキーワードから算出した重要度を管理し、キーワード群と重要度を分類処理に利用する。

【００２２】分類処理では、まず各キーワードごとに文書をまとめる単一キーワード分類処理によって、文書群を単一キーワードフォルダに格納する。複数のキーワードを持つ文書は、複数の単一キーワードフォルダに重複して分類する。

【００２３】次に、類似した文書群を含む単一キーワードフォルダについて統合判定を行い、統合可能と判断した場合には関連キーワード分類処理で統合し、関連キーワードフォルダに格納する。

【００２４】さらに、類似した文書群を含む関連キーワードフォルダについて統合判定を行い、統合可能と判断した場合には関連キーワード分類処理手段で関連キーワードフォルダの統合を繰り返す。

【００２５】単一キーワードフォルダ内あるいは関連キーワードフォルダ内について、細分類判定を行い、細分類可能な場合は単一キーワード分類処理と関連キーワード分類処理を利用して、階層的に分類する。細分類は分類停止の判断するまで再帰的に繰り返す。細分類不可能な場合は文書間関連度の判定を行い、関連度が低い文書を雑音とみなして分ける。

【００２６】各関連キーワードフォルダには、フォルダ内に格納された文書群を代表する名称を付与し、フォルダ名を付ける。

【００２７】

【作用】最終的に、既定の分類体系に依存することなく、各文書を必ず一つ以上の分類に格納し、階層的な分類体系を作成し、分類結果群を代表する名称を付与することができる。その結果、ユーザが大量の文書から所望の文書を見つけやすくすることができる。

【００２８】

【実施例】以下本発明の一実施例について説明する。

【００２９】本実施例の文書分類装置が対象とするのは、計算機上のテキストファイルであり、以下、これを文書とする。各文書には、文書の内容を代表する複数個のキーワードを付与することができ、以下、これをキーワード群とする。

【００３０】図１に本実施例の文書分類装置の構成例を示す。文書分類装置１００は、データ管理部１０１、単語検出部１０２、文書分類部１０３、分類結果出力部１０４から構成されており、一般用語を収録した単語辞書１０５、用語間の上位下位関係や同義語情報などを収録したシソーラス辞書１０６、文書を格納している文書DB１０７、各文書のキーワード群を格納しているキーワードDB１０８、出力装置のCRT１０９、入力装置のキーボード１１０、マウス１１１を持つ。

【００３１】データ管理部１０１は、文書DB１０７と、キーワードDB１０８を管理し、文書やキーワード群の入出力を行う。キーワードDB１０８には、あらかじめ人手で付与したキーワード（以下、人手付与キーワードとする）を格納することができる。人手付与キーワードは格納する必要はないが、本実施例では格納した場合を例に説明する。

【００３２】単語検出部１０２は、データ管理部１０１から文書を受け取り、単語辞書１０５を利用して形態素解析処理を行い、文書から単語を抽出する。または、文書を文字種の変化点で分割し、単語を抽出する手法や、文書中のn文字の並びに対する出現頻度を調べ、意味を持つn文字の並びを抽出する手法などを利用してもよい。抽出した単語群は、シソーラス辞書１０６を参照して、同義語展開したり、上位語への変換を行い、同義語群を生成する。抽出した単語群と、同義語群をあわせて、単語キーワードを生成する。単語キーワードをデータ管理部１０１に受け渡し、キーワードDB１０８中の各文書のキーワード群に追加登録する。

【００３３】また、単語検出部１０２は各単語キーワードの重要度を計算する。重要度とは、（１）全単語の出現頻度総数における各単語の出現頻度、（２）タイトル、見出し、段落などの文書中での出現位置に基づく。すなわち、タイトルは文書の内容を代表する可能性が高いので、タイトル中に出現した単語の重要度を高くするといった手法である。

【００３４】文書分類部１０３は、データ管理部１０１から各文書のキーワード群を受け取り、階層的な分類体系を生成し、結果を分類結果出力部１０４に受け渡す。分類処理の詳細は後述する。

【００３５】分類結果出力部１０４は、文書分類部１０３から階層的な分類体系を受け取り、インタフェース画面をCRT１０９に表示する。

【００３６】図２に単語キーワードの具体例を示し、図１の単語検出部１０２の詳細を説明する。まず、単語辞書１０５を参照して文書２０１から単語群２０２の抽出

を行う。単語群２０２は、シソーラス辞書１０６を参照して同義語や上位語に展開する。同義語群２０３は、「ブレンド」という用語が同義語の「調合」に展開された例を示していて、同義語への展開はキーワードの表記を統一するために行うものである。次に、単語群２０２と同義語群２０３から、単語キーワード２０４を生成し、これをキーワードDB１０８に格納されている文書２０１のキーワード群２０５に格納する。キーワード群２０５は、あらかじめ付与されていた人手付与キーワードに、単語キーワード２０４が追加された様子を示す。

【００３７】図３の流れ図に従い、文書分類部１０３の詳細について説明する。分類の結果、文書が格納される領域をフォルダ、フォルダに格納される複数個の文書を文書群と呼ぶ。また、「フォルダを作成する」とは、分類された文書群を格納する領域を確保することである。

【００３８】まず、各キーワードをキーワード群中に含む文書を、各キーワードごとにまとめる単一キーワード分類処理３０１を行う。次に、単一キーワード分類処理３０１によって作成された単一キーワードフォルダについて、類似した文書群を含む単一キーワードフォルダの組を統合する関連キーワード分類処理３０２を行う。ステップ３０３では、関連キーワード分類処理３０２によって作成された関連キーワードフォルダについて、類似した文書群を含む関連キーワードフォルダの組を統合できるかどうかの判定を行い、統合が可能な間は関連キーワードフォルダの統合を繰り返す。さらに、作成された単一キーワードフォルダや関連キーワードフォルダ内について、細分類できるかどうかを調べ（ステップ３０４）、細分類可能な間は階層的に分類を繰り返す（ステップ３０５）。ステップ３０５は、すべてのフォルダ内について階層的な分類を行う。

【００３９】次に、図３の各処理の詳細を説明する。

【００４０】まず、図４の流れ図を用いて、図３の単一キーワード分類処理３０１の詳細を説明する。以下、単一キーワード分類処理３０１で作成するフォルダを単一キーワードフォルダと呼ぶ。

【００４１】ステップ４０１では、キーワードを管理するテーブルの初期化を行う。キーワードを管理するテーブルをキーワードテーブルと呼ぶ。ステップ４０２では、図１のキーワードDB１０８から一文書のキーワード群を読み出し、キーワードテーブルに各キーワードを登録する。すべての文書について、キーワードの登録を繰り返す（ステップ４０３）。ステップ４０４では、各キーワードについて、キーワードをキーワード群に含む文書数を数える。文書数が１であるかどうかの判定を行い（ステップ４０５）、１のときはキーワードを対象外とする（ステップ４０６）。一文書にしか含まれないキーワードは、分類時に他の文書との関連性を判断する材料にならないためである。文書数が１ではないときは、キーワードと同一の名前の単一キーワードフォルダを作成し、キーワードをキーワード群中に含む文書群を単一キーワードフォルダに格納し、さらに単一キーワードフォルダ名をフォルダを管理するテーブルに登録する（ステップ４０７）。フォルダを管理するテーブルをフォルダテーブルと呼ぶ。ステップ４０８では、全キーワードについて、上述した単一キーワードフォルダの作成処理を繰り返す。

【００４２】次に、上述したキーワードテーブルについて図５を用いて説明する。キーワードテーブル５０１の各エントリは、キーワードリスト５０２を指している。エントリは、キーワードを入力とするハッシュ関数の値で決定する。キーワードリスト５０２は、キーワード５０３、同一ハッシュ値のキーワードリストへのポインタ５０４、文書識別番号リストへのポインタ５０５の組である。ポインタ５０４は、同一ハッシュ値のキーワードリスト５０６を指していて、キーワードリスト５０２と同一ハッシュ値のキーワードリスト５０６のキーワードは同じハッシュ値になることを示す。ポインタ５０５は、文書識別番号リスト５０７を指していて、キーワード５０３を含む文書群が連結されている。文書識別番号リスト５０７は、文書を識別するための番号５０８、次文書識別番号リストへのポインタ５０９の組である。各キーワードを含む文書の数は、キーワードテーブル５０１の文書識別番号リスト５０５をたどれば得られる。

【００４３】図６に示したキーワードテーブルの具体例では、キーワードテーブル６００の７番のエントリ６０１にキーワードリストが連結されている。ポインタ６０３は、「ペット」というキーワード６０２のキーワードリストに、「新種」というキーワード６０９のキーワードリストが連結していることを示し、ポインタ６１０はキーワードリストの連結はそれ以上はないことを示している。ポインタ６０４は、「ペット」というキーワードを含む一つの文書識別番号リストである文書識別番号１（６０５）を指し、さらにポインタ６０６は文書識別番号２（６０７）を指している。ポインタ６０８は、それ以上「ペット」というキーワードを含む文書群はないことを示す。同様に、ポインタ６１１は「新種」というキーワードを含む文書識別番号リストの文書識別番号３（６１２）を指している。ポインタ６１３は、さらに文書識別番号リストが連結されていることを示す。

【００４４】次に、フォルダテーブルについて図７を用いて説明する。フォルダテーブル７００は、フォルダ名７０１、フォルダに格納されている文書数７０２、分類終了フラグ７０３、文書識別番号リスト７０４の組である。分類終了フラグ７０３は分類を行うか否かを記憶するフラグで、詳細については後述する。文書識別番号リスト７０４は、図５のキーワードテーブル５０１の文書識別番号リスト５０７と等しい。フォルダテーブル７００の０番のエントリ７０５は、「ペット」というフォルダがあり、その中には文書が２件格納されていて、フォ

ルダの分類はOFF（終了していない）で、文書識別番号のリストが連結していることを示す。

【００４５】図８に示す単一キーワード分類処理の具体例を用いて、図３の単一キーワード分類処理３０１を説明する。文書８０１は図１の文書DB１０７に格納されていて、文書群８０２を形成する。各文書にはキーワード群８０３が付与されていて、キーワード群８０３は個々のキーワード８０４の集合である。８０３１は文書１のキーワード群で、「犬」、「動物」、「ペット」というキーワードを含む。８０３２は文書２のキーワード群で、「猫」、「動物」、「ペット」、「ねずみ」というキーワードを含む。８０３３は文書３のキーワード群で、「新種」、「ねずみ」、「動物」、「ABC国」というキーワードを含む。８０３４は文書４のキーワード群で、「ねずみ」、「キャラクター商品」、「ABC国」というキーワードを含む。その他の文書５、文書６、文書７、文書８も同様にキーワード群を持つものとする。文書群８０２に対して単一キーワード分類処理を行うと、単一キーワードフォルダ８００群に展開される。単一キーワードフォルダは、「犬」、「猫」、「キャラクター商品」、「新種」、「ABC国」、「ペット」、「ねずみ」、「動物」、というキーワードと同一の名前で作成されている。各単一キーワードフォルダには、文書集合８０５が格納されている。

【００４６】例えば、８０５１は単一キーワードフォルダ「犬」の文書集合で、文書１が含まれていることを示す。８０５２は、単一キーワードフォルダ「ペット」の文書集合で、文書１と文書２が格納されている。８０５３は単一キーワードフォルダ「ねずみ」の文書集合で、文書２、文書３、文書４が格納されている。８０５４は単一キーワードフォルダ「動物」の文書集合で、文書１、文書２、文書３が格納されている。８０５５は単一キーワードフォルダ「ABC国」の文書集合で、文書３、文書４が格納されている。８０５６は単一キーワードフォルダ「新種」の文書集合で、文書３、文書５、文書６、文書７、文書８が格納されていることを示す。

【００４７】次に、図９の流れ図に従い、図３の関連キーワード分類処理３０２の詳細を説明する。以下、関連キーワード分類処理３０２で作成するフォルダを関連キーワードフォルダと呼ぶ。

【００４８】ステップ９０１では、フォルダ間の一致文書数を管理するテーブルの初期化を行う。フォルダ間の一致文書数を管理するテーブルをフォルダ間一致文書数管理テーブルと呼ぶ。二つの単一キーワードフォルダ間での一致文書数を数えて、フォルダ間一致文書数管理テーブルに登録する（ステップ９０２）。ステップ９０３では、すべての単一キーワードフォルダ間の一致文書数をフォルダ間一致文書数管理テーブルに登録する。ステップ９０４では、フォルダ間一致文書数管理テーブルを一致文書数で降順に配列し、テーブルの先頭すなわちフォルダ間の一致文書数が最大となった単一キーワードフォルダの組の統合が可能であるかを調べる（ステップ９０５）。ステップ９０５の詳細は後述する。

【００４９】統合が可能である場合は、関連キーワードフォルダを作成し、統合する二つの単一キーワードフォルダの文書群を格納し、関連キーワードフォルダ名を図７のフォルダテーブル７００に登録し、統合した二つの単一キーワードフォルダをフォルダテーブル７００から削除する（ステップ９０７）。関連キーワードフォルダの名称は、統合した二つの単一キーワードフォルダ名を列挙したものである。このとき、列挙する順番は文書数の多い単一キーワードフォルダ名から並べ、関連キーワードフォルダ内の文書群がどのようなキーワードを多く含むかを明示する。さらに、ステップ９０８でフォルダ間一致文書数管理テーブルに統合が終了したことを書き込み、フォルダ間一致文書数管理テーブルの更新を行う。

【００５０】統合が不可能である場合や、統合が終了した場合には、統合処理を行っていない単一キーワードフォルダの組について、フォルダ間一致文書数管理テーブルの先頭から終りまで、統合処理を繰り返す（ステップ９０６）。

【００５１】上述したフォルダ間一致文書数管理テーブルを図１０に示し、詳細を説明する。フォルダ間一致文書数管理テーブル１００１は、一致文書数を調べる二つのフォルダの識別番号１００２、１００３、一致文書数１００４、統合済フラグ１００５の組である。図１０の例では、フォルダ識別番号０番と１番の一致文書数が５件あったことを示している。統合済フラグ１００５は、「フォルダの統合が済んでいるか否か」や「フォルダの統合が不可能である」ことを記憶するフラグで、「済」は統合が済んでいることを示し、「不可」は統合が不可能であることを示している。

【００５２】次に、図１１にフォルダ間一致文書数管理テーブルの具体例を示す。すべての単一キーワードフォルダの識別番号を１１０１と識別番号１１０２に列挙し、二つのフォルダ間の一致文書数を調べて登録したもので、例えば、文書０と文書４の一致文書数１１０３は８件あったことを示す。一致文書数は、図７のフォルダテーブル７００の文書識別番号リスト７０４をたどり、比較することで求められる。

【００５３】図１２にフォルダ間一致文書数管理テーブルの更新処理の具体例を示し、これを用いて詳細を説明する。フォルダ間一致文書数管理テーブル１２００は、図９のステップ９０４まで処理が終了した状態を示す。フォルダ間一致文書数管理テーブルの先頭、すなわちエントリ０番からフォルダの統合処理を開始し、フォルダ識別番号０番と１番の統合が可能と判断されると統合を行い（ステップ９０５）、統合済フラグ１２０５に「済」と書き込む（ステップ９０８）。フォルダ識別番

号０番と１番を統合すると、０番と１番に関係する１２０６～１２１３を「不可」にする。次に、フォルダ間一致文書数テーブルの次のエントリ、すなわちエントリ１番について統合処理を行う（ステップ９０６）。このとき、統合済フラグが「不可」ではないことを確認する。「不可」ならばエントリ１番の単一キーワードフォルダの組の一方は、すでに統合処理が終了しているので統合はできない。図１２では、統合済フラグ１２１４は「不可」ではないので、統合を行うことができる。本実施例では、フォルダ間一致文書数管理テーブルの先頭から順番に処理を行うので、「済」のエントリが処理中のエントリよりも後ろのエントリに現われることはない。

【００５４】次に、図１３の流れ図を用いて、図９のステップ９０５のフォルダの統合判定処理の詳細を説明する。

【００５５】統合するフォルダの組は、「一致文書数が最大になること」を条件として決定する。統合すべきかどうかは、統合した結果が有効な分類になるかどうかを調べる必要があり、統合前と統合後を比較して判定する。本実施例では図１３に示すフォルダ内文書の距離計算処理１３００を適用する。

【００５６】図１３の１３０１は、フォルダ内文書の距離計算処理１３００で一時的に用いる作業用キーワードテーブルの初期化を行っている。作業用テーブルのデータ構造は、図５のキーワードテーブル５０１と同様とする。フォルダ内に格納されている文書に対して、キーワードの登録を行い（ステップ１３０２）、すべてのフォルダ内文書について繰り返す（ステップ１３０３）。キーワード数の計数用にキーワード数pを初期化する（ステップ１３０４）。キーワードを含む文書数を数えて（ステップ１３０５）、文書数が１のキーワードは作業用キーワードテーブルから削除し（ステップ１３０８）、１以上のときはキーワード数pを１ずつ増やす（ステップ１３０７）。すべてのキーワードについてステップ１３０５以下の処理を繰り返し（ステップ１３０９）、フォルダ内の文書群に含まれるキーワードのうち、二つ以上の文書に含まれるキーワードの選定が終了する。次に、図１４のステップ１４００に進む。

【００５７】図１４の１４０１では、ワードベクトルを管理するテーブルの初期化を行う。ワードベクトルWiとは、文書iにおける「キーワードの出現頻度とキーワードの重要度の積」の並びであり、具体的には次のように表現できる。

【００５８】ワードベクトルWi = (F1*V1, F2*V2, ..., Fj*Vj, ..., Fp*Vp)

(iは文書識別番号、1<=j<=p,jはキーワード識別番号、pはキーワード数、Fjはキーワードjの出現頻度、Vjはキーワードjの重要度)

重要度は、図１の単語検出部１０２で付与したもので、値が大きいものほど重要度が高い。ワードベクトルを管理するテーブルをワードベクトルテーブルと呼ぶ。

ステップ１４０２では、各キーワードの出現頻度と単語検出部１０２で付与した各キーワードの重要度の積を計算し、ワードベクトルテーブルに登録し、各文書に関して繰り返す（ステップ１４０３）。さらに、ステップ１４０４では、各文書のワードベクトルの平均ベクトルを求める。本実施例では次式で定義する。

【００５９】平均ベクトルWa = ΣWi / n

(1<=i<=n, iは文書識別番号、nは文書数)

次に、ステップ１４０５では、各文書のワードベクトルWi (1<=i<=文書数) と平均ベクトルWaとの距離を計算する。ベクトル間の距離とは、ベクトルの近さを判断するもので、本実施例ではベクトル間の距離を次式で定義する。文書Di、Djのワードベクトルをそれぞれ Wi、Wjとし、WiとWjのなす角度を θ とし、Di、Dj間の距離をd(Di, Dj)とする。

【００６０】

d(Di,Dj) = 1 - Wi・Wj / |Wi|*|Wj| = 1 - cos θ

（ただし、・は内積、*は積、|Wi|はWiの大きさ）

d(Di,Dj)は0<=d(Di,Dj)<=1の範囲の値であり、ベクトル間の距離が近いほど値は小さくなり、一致する場合は0になる。

【００６１】すべての文書について、ワードベクトルと平均ベクトルとの距離計算を繰り返す（ステップ１４０６）。次にステップ１４０７で、すべての平均ベクトルと各文書との距離から距離分布を求める。距離分布とは、（１）平均距離、（２）分散とし、次式で定義する。

【００６２】平均距離da = Σdi / n

分散 σ = Σ((di - da) * (di - da)) / (n-1)

(1<=i<=n, iは文書識別番号、nは文書数、diは文書識別番号iの文書と平均ベクトルとの距離)

統合前の二つの単一キーワードフォルダについて、別々に調べた距離分布の平均値と、統合後の関連キーワードフォルダの距離分布を比較することで、統合の可否を判断する。本実施例では、（１）、（２）を具体的に次式で定義する。

【００６３】（１）|d2-d1| >= Td

(d1は統合前の二つの単一キーワードフォルダの平均距離の平均値、d2統合後の平均距離、|x|はxの絶対値、Tdはしきい値)

（２）σ2/σ1 >= Tσ

（σ1は統合前の二つの単一キーワードフォルダの平均分散値、σ2は統合後の分散値、Tσはしきい値）

（１）または（２）の条件が満たされるときに、統合不可と判断する。しきい値Td、Tσは、初期実験で数種類を決定しておき、フォルダ内の文書群の距離分布の状況に適したものを選択する。

【００６４】図１５のワードベクトルテーブルの具体例を用いて、詳細を説明する。ワードベクトルテーブル１

１５

５００は、縦軸が文書識別番号、横軸がキーワード識別番号の２次元テーブルである。例えばエントリ１５０３は、文書識別番号３番の文書において、キーワード識別番号３のキーワードの出現頻度と重要度の積が２であることを表している。図１５を例にとり、ワードベクトルと平均ベクトルとの距離の計算例を示す。説明を簡略化するために、各キーワードの重要度はすべて１とし、文書数は４、キーワードは識別番号４までを対象としたときの、文書識別番号１番のワードベクトルと平均ベクトルとの距離を計算する。

【００６５】W1 = (3, 2, 1, 1)　　*W2 = (1,13, 2, 0)
W3 = (1, 1, 8, 0)
W4 = (0, 1, 3, 2)
平均ベクトルWa = ΣWi / 4
= (5, 17, 14, 3) / 4
= (1.3, 4.3, 3.5, 0.8)
(少数第２位を四捨五入)
文書識別番号１のワードベクトルと平均ベクトルWaとの距離d1

$$d1(D1, Wa) = 1 - W1 \cdot Wa / |W1| * |Wa|$$
$$= 1 - (3, 2, 1, 1) \cdot (1.3, 4.2, 3.5, 0.8) / 3.9 * 38$$
$$= 1 - 48.1 / 148.2$$
$$= 1 - 0.32$$
$$= 0.68$$

図１６に関連キーワード分類処理の具体例を示す。単一キーワードフォルダ「ペット」１６０１と単一キーワードフォルダ「動物」１６０２とを統合して関連キーワードフォルダ「動物とペット」１６０５を作り、単一キーワードフォルダ「ABC国」１６０３と単一キーワードフォルダ「ねずみ」１６０４を統合して関連キーワードフォルダ「ねずみとABC国」を作成している（図３ステップ３０２）。さらに、関連キーワード分類処理を繰り返し（図３ステップ３０３）、関連キーワードフォルダ「動物とペット」１６０５と関連キーワードフォルダ「ねずみとABC国」１６０６を統合し、関連キーワードフォルダ「動物とねずみとペットとABC国」１６０７を作成している。１６０７のフォルダ名は、「動物」、「ねずみ」、「ペット」、「ABC国」の順に、各キーワードに関連する文書が多く格納されていることを示す。

【００６６】関連キーワード分類処理が終了すると、分類体系の第一階層が生成される。

【００６７】次に、細分類処理について説明する。細分類とはフォルダ内を階層的に分類することで、上位フォルダの作成に利用していないキーワードを用いて分類を行う。例えば、図１６の関連キーワードフォルダ「動物とねずみとペットとABC国」１６０７内を細分類する場合には、キーワード「動物」、「ねずみ」、「ペット」、「ABC国」以外のキーワードを利用して分類を行う。

【００６８】図１７に細分類の流れ図を示し、図３のステップ３０４の詳細を説明する。ステップ１７０１ではフォルダの種類を判別する。

【００６９】関連キーワードフォルダの場合は、類似度の高い、複数の単一キーワードフォルダが統合された結果なので、フォルダ内をさらに階層的に分類する。分類は、図３の流れ図に従って、単一キーワード分類処理３０１、関連キーワード分類処理３０２、関連キーワードフォルダの統合処理３０３、細分類３０４～３０６を再※帰的に繰り返す。

【００７０】単一キーワードフォルダの場合は、一つのキーワードに引き付けられた文書群が格納されているので、集合としてのまとまりがあるかどうかの保証がない。そこで、図１３のフォルダ内の文書間の距離計算を行って（ステップ１３００）、各文書と平均ベクトルとの距離の分散値を求める（ステップ１７０２）。

【００７１】分散値としきい値Tσを比較し（ステップ１７０２）、Tσ以上ならばフォルダ内を階層的に分類する価値はないとみなし、さらに平均距離を調べる（ステップ１７０３）。平均距離がしきい値Td以上の文書は、関連性の薄い雑音文書であると判定し、単一キーワード内の雑音フォルダに格納する（ステップ１７０４）。雑音フォルダは、雑音と判定された文書を格納するためのフォルダで、雑音文書が存在したフォルダ内にだけ作成する。フォルダ内のすべての文書について、平均距離計算を行い（ステップ１７０５）、雑音文書を雑音フォルダに格納し、図７のフォルダテーブル７００の分類終了フラグ７０３に分類終了と書き込む（ステップ１７０６）。

【００７２】分散値がTσ以下の場合は、単一キーワードフォルダ内を細分類可能と判断し、図３の流れ図に従って、単一キーワード分類処理３０１、関連キーワード分類処理３０２、関連キーワードフォルダの統合処理３０３、細分類３０４～３０６を再帰的に繰り返す。

【００７３】図１８の細分類の具体例では、関連キーワードフォルダ「猫と魚」１８０１は、単一キーワードフォルダ「キャットフード」１８０４と、関連キーワードフォルダ「釣りと海」１８０５の二つのフォルダに階層的に分類された例である。また、単一キーワードフォルダ「犬」１８０２は、単一キーワードフォルダ「柴犬」１８０６と関連キーワードフォルダ「えさと散歩」１８０７の二つのフォルダに階層的に分類された例である。単一キーワードフォルダ「新種」１８０３は、雑音文書

１８０９が雑音文書フォルダ１８０８に分けられた例である。

【００７４】細分類によって、分類体系の第二階層以下を作成したことになる。

【００７５】これまでは文書分類部の流れを説明した。ここで、分類によって作成されるフォルダの階層構造の記憶方法について、図１９の分類階層管理テーブル１９００を用いて説明する。

【００７６】分類階層管理テーブル１９００の各エントリは、フォルダ情報リストを指している。フォルダ情報リスト１９０１は、フォルダ名１９０２、文書識別番号リスト１９０３、文書数１９０４、兄弟フォルダ情報リストへのポインタ１９０５、子フォルダ情報リストへのポインタ１９０６、親フォルダ情報リストへのポインタ１９０７の組である。フォルダ名１９０２は図７のフォルダテーブル７００のフォルダ名７０１と一致し、文書数１９０４は７０２と一致する。文書識別番号リスト１９０３は、各フォルダに格納されている文書識別番号リスト１９０８へのポインタであり、１９０８は文書識別番号１９０９と次文書識別番号リストへのポインタ１９１０の組である。文書識別番号リスト１９０３は、図７のフォルダテーブル７００の文書識別番号リスト７０４と一致する。兄弟フォルダ情報リストへのポインタ１９０５は、フォルダ情報リスト１９０１と同じ上位フォルダを持ち、同階層に位置するフォルダ情報リストへのポインタである。子フォルダ情報リストへのポインタ１９０６は、フォルダ情報リスト１９０１の一つの下位フォルダ情報リストへのポインタである。親フォルダ情報リストへのポインタ１９０７は、上位フォルダ情報リストへのポインタである。

【００７７】図７のフォルダテーブル７００に作成したフォルダの情報を書き込むと同時に、分類階層管理テーブル１９００にもフォルダテーブル７００の内容を複写する。細分類によって、第二階層以降のフォルダを作成すると、分類階層管理テーブル１９００の子フォルダ情報リストへのポインタ１９０６、兄弟フォルダ情報リストへのポインタ１９０５、親フォルダ情報リストへのポインタ１９０７の更新を行う。

【００７８】図１９の分類階層管理テーブル１９００を用いて、具体的に階層構造の記憶方式を説明する。まず、新規作成したフォルダをフォルダ情報リスト１９０１に登録する。さらに、フォルダ内を細分類して二つのフォルダに分類したとすると、子フォルダ情報リストへのポインタ１９０６に一つの子フォルダ情報リスト１９１１を登録し、１９１１の兄弟フォルダ情報リストへのポインタ１９１２に二つ目の子フォルダ情報リスト１９１５を登録し、それ以上はフォルダはないので、１９１６は連結がないことを示す。フォルダ情報リスト１９０１は１９１１と１９１５の親フォルダ情報リストに相当するので、親フォルダ情報リストへのポインタ１９１４、１９１８は１９０１を指している。フォルダ情報リスト１９０１の子フォルダ情報リスト１９１１、１９１５は、以下に階層的な分類はないので、１９１３、１９１８は連結がないことを示す。フォルダ情報リスト１９０１は第一階層のフォルダで、それ以上の階層や同一の階層に位置するフォルダはないので、１９０５、１９０７は連結がないことを示す。また、フォルダ情報リスト１９０１は、二つの文書を持ち、１９０３は一つ目の文書識別番号リスト１９０８を、１９０９は二つ目の文書識別番号リスト１９１０を指している。

【００７９】図１の文書分類部１０３で生成された分類体系は、具体的には図２０のような階層構造に展開され、第一階層には、関連キーワードフォルダ「猫と魚」１８０１、単一キーワードフォルダ「犬」１８０２、単一キーワードフォルダ「新種」が、第二階層フォルダには、１８０１の下位に単一キーワードフォルダ「キャットフード」１８０４、関連キーワードフォルダ「釣りと海」１８０５があり、１８０２の下位に単一キーワードフォルダ「柴犬」１８０６、関連キーワードフォルダ「えさと散歩」１８０７があり、１８０３の下位には雑音文書が雑音文書フォルダに分離されている。

【００８０】図１の分類結果出力部１０４は、上記分類体系を文書分類部１０３から受け取ると、インタフェース画面を図２１に示すように作成し、CRT１０９に出力する。図２１の２１０１は、分類体系の上位三階層が表示されていて、２１１３は第一階層、２１１４は第二階層２１１５が第三階層を示していて、各階層のフォルダ名が縦方向に表示されている。図２１は、第一階層「猫、魚」を選択し、第二階層「キャットフード」を選択した結果、第三階層に文書群が表示され、文書１５を選択した様子である。２１１２は文書１５の内容、文書の作成された日２１１６やフォルダ内における文書１５の得点情報２１１７を表示している。ユーザは図１のマウス１１１で、興味のあるフォルダ名を選択し、自由に内容を参照することができる。また、興味のないフォルダは読み飛ばすことができ、参照すべき文書量が削減できる。

【００８１】図１の文書DB１０７に新規文書の到着や、古い文書の削除が行われる場合には、文書分類部１０３が改めて分類しなおすことで、新規情報の入手にも対応できる。

【００８２】次に、第２の実施例を説明する。第２の実施例は、図１に示した第１の実施例の文書分類装置１００における分類結果出力部１０４に、ユーザの意見を反映した分類結果を構築するための分類指定手段を設けた文書分類装置に関する。

【００８３】第１の実施例の文書分類装置は既定の分類体系にとらわれることなく文書を自動的に分類するために、ユーザの意向と異なる分類結果を生成してしまうことがある。そこで図２２のように、図１の分類結果出力

部１０４に分類指定部２２０１を加えた、文書分類装置２２００の構成をとり、ユーザの意見を分類結果に反映させる手段を設ける。分類指定部２２０１はフォルダ数指定インタフェース２５００と分類体系構築補助インタフェース２７００という二つの画面から構成される。

【００８４】ユーザがキーボード１１０、マウス１１１といった入力装置を用いて分類したい文書群を指定し、分類処理の実行を指示すると、文書分類装置２２００が起動され、図２３の流れ図に基づく処理が施される。まずステップ２３０１としてデータ管理部１０１が文書ＤＢ１０７にユーザの指定した文書群を格納する。

【００８５】続くステップ２３０２では、単語検出部１０２が文書群から単語キーワードを検出し、単語辞書１０５に格納する。こうして分類処理を行うためのデータが用意できると、ステップ２３０３で文書分類部１０３が図３の流れ図に従って分類体系を生成し、文書を分類する。分類結果出力部１０４はこの分類結果を図２１のようなインタフェース画面に表示して、ユーザに提示する（ステップ２３０４）。ここまでは第１の実施例と同一の処理ステップである。さらに、分類結果を参照したユーザから分類指定部２２０１に対する指示を確認し（ステップ２３０５）、指示がない場合には終了する。指示がある場合にはステップ２３０６で指示内容を解釈し、フォルダ数指定インタフェース２５００への指示の場合はステップ２３０７の再分類処理を施し、分類体系構築補助インタフェース２７００への指示の場合はステップ２３０８の再分類処理を適用する。フォルダ数指定インタフェース２５００および分類体系構築補助インタフェース２７００については後述する。そして再びステップ２３０４に戻り、再分類の結果をユーザに提示する。ユーザが分類指定部２２０１に対して指示するケースとしては、分類結果がユーザの意向に合わない場合などが挙げられる。

【００８６】次に、分類指定部２２０１の提供する二つの入力画面、フォルダ数指定インタフェース２５００および分類体系構築補助インタフェース２７００について説明する。ここでは、各インタフェースから取り込まれる再分類に関する指示情報と再分類処理２３０７および２３０８の詳細について述べる。

【００８７】はじめにフォルダ数指定インタフェース２５００について述べる。

【００８８】図２４は分類結果出力部１０４によって作成された、約一千件の「コンピュータ」関係の文書群の分類結果を示すインタフェース画面である。これは図２１と同種の出力画面であり、第１の実施例と同様に文書分類部１０３で作成された分類結果を読み出して作成される。具体的には図１９の分類階層管理テーブル１９００からフォルダ名１９０２、文書群１９０３、文書数１９０４、フォルダの階層関係（１９０５、１９０６、１９０７）を読み出すことによって、図２４の画面に表示

する情報を得ている。

【００８９】このように一つの分野の文書群であっても内容が多岐に渡れば詳細に分類することができ、数十から数百ものフォルダが生成される。分類結果として第一階層に数十個のフォルダが生成され、各フォルダの下位に同等数のフォルダが生成された場合を例に、本図では分類体系の上位三階層を表示している。フォルダ「パソコン、発売、販売、ソフト」２４０５およびフォルダ「発売、価格、販売、見込」２４０６はユーザが参照しようとして選択した状態を反転して示している。２４０１は第一階層に生成された複数個のフォルダの名前を縦方向に列挙し、２４０２はユーザから選択された第一階層２４０１のフォルダ「パソコン、発売、販売、ソフト」２４０５の下位（第二）階層のフォルダ名を、２４０３は選択された第二階層２４０２のフォルダ「発売、価格、販売、見込」２４０６の下位（第三）階層のフォルダ名を縦方向に表示している。本図では第一階層（top class）２４０１のフォルダ名が６個しか見えないが、実際にはスクロールバー２４０４を用いて画面をスクロールさせることで数十個のフォルダ名を参照できる。

【００９０】なお第一階層内をさらに細分化し第二階層以下を作成する細分類処理（図１７）について、第１の実施例では「上位フォルダの作成に利用していないキーワードを用いて分類を行う」場合を仮定したが、本例では上位フォルダの作成に利用したキーワードも用いて分類する場合について取り挙げている。そのため、上位階層のフォルダの作成に利用されたキーワードが下位階層のフォルダにも出現することがあり、例えば第一階層（top class）２４０１のフォルダ「パソコン、発売、販売、ソフト」２４０５に含まれているキーワード「パソコン」、「発売」、「販売」、「ソフト」は第二階層２４０２のフォルダ「発売、価格、販売、見込」２４０６とフォルダ「パソコン、開発、シリーズ、新製品」２４０７、フォルダ「ソフト、東京、複雑、成功」２４０８にも含まれている。

【００９１】この分類結果では一階層のフォルダ数が多く、所望のフォルダを見つけにくい。一方、一階層のフォルダ数を少なくし階層を深くして細分類すると、所望の文書を見つけるまで手間がかかる。分類結果として適切なフォルダの数やフォルダの大きさは、分類対象の文書数や文書の内容の均質さによって異なってくる。さらに、分類結果の適否はこれを参照するユーザの視点によって異なることから、予め適切なフォルダ数や平均文書数を設定することは難しい。そこで、フォルダ数指定インタフェース２５００は文書分類部１０３によって生成される分類結果のフォルダ数やフォルダの大きさを、ユーザによって指定できる環境を提供するという役割を担う。

【００９２】図２５に示すフォルダ数指定インタフェー

ス２５００では、フォルダ数やフォルダの平均文書数といった分類結果に関する情報をユーザに提示し、好みに応じて適切なフォルダ数や平均文書数にまとめ直すための指示を受け取れるようにしている。一般的にフォルダ数を少なくすれば平均文書数は多くなるというように両者は連動する関係にあるが、ユーザが分類結果を評価する基準としてどちらを用いてもよいようにする。２５０１は一階層に生成されたフォルダ数を、２５０２は平均文書数を示す。フォルダ数操作バー２５０３はフォルダ数の増減を、平均文書数操作バー２５０４は平均文書数の増減をユーザが指定するためのグラフィカルユーザインタフェース（GUI）である。２５０５は文書数に応じた大きさに表現した円状の図形をフォルダと見立て、分類体系全体の状況を表示している。分類指定部２２０１はフォルダ数指定インタフェースを作成するために、分類結果出力部１０４から第一階層２４０１に関する情報を読み出し、フォルダ数を調べて２５０１に書き込み、各フォルダの文書数を調べてその平均値を２５０２に書き込み、各フォルダは文書数に比例した値を半径とする円状の図形として画面２５０５を作成する。

【００９３】図２５はユーザがフォルダ数操作バー２５０３を操作してフォルダ数の減少を指示した様子を示している。ユーザからの指示に従って再分類処理２３０７が実行されると、新たな分類結果が生成されフォルダ数指定インタフェースは２５００から２５１０のように変わる。再分類処理２３０７の詳細は後述する。

【００９４】再分類前にはフォルダ数２５０１は９６個であったが、ユーザの指示を反映して、フォルダ数２５１１は３０個に減少している。このように、ユーザが自分の参照しやすいレベルに分類結果を調節することが可能となる。

【００９５】次に、分類指定部２２０１がフォルダ数指定インタフェース２５００から取り込まれたユーザの指示に基づいて行う、再分類処理２３０７の詳細を図２６を用いて説明する。

【００９６】まず、分類指定部２２０１はステップ２６０１としてユーザからの指示内容を解釈する。「フォルダ数の減少」あるいは「平均文書数の増加」を指示するものでなければ、ステップ２６０２として分類処理２３０３における分類結果を見直し、よりフォルダ数が多く、平均文書数が少ない分類結果を再選択することを文書分類部１０３に指示する。

【００９７】ここで文書分類部１０３が行う、分類結果の再選択処理２６０２について説明を加える。文書分類部１０３の行う分類処理２３０３は図３の流れ図に基づき、関連キーワード処理３０２を繰り返し適用して分類結果となるフォルダを生成し、これらのフォルダに文書を分類する。関連キーワード処理３０２とは、図９で示した流れ図に従い、関連のありそうなフォルダの組を統合することによって分類結果となるフォルダを作り出す。フォルダに関する情報を記録しておくフォルダテーブル７００（図７）は、関連キーワード分類処理３０２が繰り返されるたびに更新され、関連キーワード処理３０２が終了（ステップ３０３）した時点の情報が分類結果のフォルダとして採用される。

【００９８】すなわち、関連キーワード分類処理３０２が繰り返されるたびにフォルダが統合されて、分類結果全体としてみるとフォルダ数が減少し、平均文書数が増加する。そこで、関連キーワード分類処理３０２のたびにフォルダテーブル７００の情報を中間結果的なフォルダとして記録しておけば、後から分類結果よりも多いフォルダ数で、平均文書数の少ないフォルダを再選択することが可能となる。例えば第１の実施例の図１６のフォルダ「動物とねずみとペットとABC」１６０７が分類結果のフォルダの一つと仮定すると、中間結果にはフォルダ「動物とペット」１６０５、フォルダ「ねずみとABC国」１６０６が該当する。中間結果は関連キーワード処理３０２を繰り返した回数分だけ存在するので、ユーザの指示した操作バーの増減レベルに応じて「より少ないフォルダ数」で「より多い平均文書数」の中間結果を選択することによって、ユーザからの指示に対応する。

【００９９】一方ステップ２６０１で、ユーザから「フォルダ数の減少」および「平均文書数の増加」が指示された場合は、分類指定部２２０１が文書分類部１０３に対して「より少ないフォルダ数で、より平均文書数の多い分類結果を作り直すこと」を指示する（ステップ２６０３）。これに対して文書分類部１０３では、分類結果のフォルダをさらに統合することによって「より少ないフォルダ数で、より平均文書数の多い分類結果」を作る。分類処理２３０３で分類結果となるフォルダが生成されるのは、図３の流れ図のステップ３０３で関連キーワード分類処理が終了したと判断される場合、すなわち図９のステップ９０５でフォルダの統合がこれ以上不可能であると判定された場合である。そこでステップ２６０３では、ステップ９０５の統合判定条件となるしきい値TdおよびTσを調整して、フォルダの統合がさらに可能となるような設定を行う。第１の実施例の図１７の説明で述べたように、しきい値Tdは平均ベクトルとの距離がTd以上に離れている場合に統合不適と判断する指標なので、Tdをより大きい値に設定し直す。しきい値Tσは平均ベクトルとの距離の分散値がTσ以上に大きい場合に統合不適と判断する指標なので、Tσをより大きい値に設定し直す。具体的な値はユーザの指示した操作バーの増減レベルに応じて、文書分類部１０３が決定する。これら準備の後、関連キーワード処理３０２を適用してさらなるフォルダの統合を行い、可能な限りフォルダの統合を繰り返す（ステップ３０３）。これによって先にユーザに提示した分類結果よりもフォルダ数が少なく、平均文書数が多い分類結果を生成できるので、この結果から分類結果の再選択処理２６０２を行う。

【０１００】次に、分類指定部２２０１の提供するもう一つの入力画面である、分類体系構築補助インタフェース２７００の詳細を述べる。

【０１０１】文書分類部１０３はシソーラスなどの情報を用いずに、自動的にフォルダを階層化するため、一般的な上位語下位語の概念と矛盾する階層関係を生成する場合がある。例えば、図２０のフォルダ「犬」１８０２とフォルダ「柴犬」１８０６は正しい上下関係にあるが、逆転した場合には概念的に矛盾することになる。

【０１０２】この解決策として、シソーラス辞書１０６を用いて上位語下位語として不適当な関係となるフォルダの生成を禁止する方法が考えられる。しかし、本発明の文書分類装置が生成するフォルダ名はキーワードを統合した形式、フォルダ「猫と魚」１８０１のようになることが多い。そこで前例の「犬」と「柴犬」のように、１対１のキーワードの上下関係を調べるのでは対応し切れず、複数個対複数個のキーワードの上下関係を考慮しなければならない。このとき、例えばキーワードA、B、C、Dから上位フォルダ「A、B」、下位フォルダ「C、D」が作られたとすると、「キーワードAとCは上位語下位語として適当であるが、キーワードBとDは逆転関係にある」という場合には適否判断が行えないという問題が残り、これでは不十分である。

【０１０３】分類体系構築補助インタフェース２７００は、適切な階層構造を作るための補助情報をユーザから取り込み、これを用いて文書分類部１０３が分類処理を行えるようにするものである。例えば特許の明細書ならば「発明の名称」、「特許請求の範囲」といった特定の項目と各項目に書くべき内容が定められているが、これらが補助情報に相当する。各項目に書かれる文章の内容は明細書ごとに異なるが、専門性や一般性の度合には共通点があり、項目ごとにその度合が決まっているものと考えられる。例えば明細書の「発明の名称」や「発明の属する技術分野」といった項目には発明の前提条件や背景が書かれるので、他の項目に比べて一般性が高い。また、「課題を解決するための手段」や「発明の実施の形態」などの項目には発明の内容が記載されることから、専門性が高くなる。上位語となるキーワードは一般性が高く、下位語となるキーワードは下位に位置するほど専門性が高い。

【０１０４】したがって、一般性の高い内容の項目に出現するキーワードを上位階層のフォルダに、専門性の高い内容の項目に出現するキーワードを下位階層のフォルダの作成に利用することで適切な階層構造を構築しやすくなる。具体的には、一般性の高い「発明の名称」や「発明の属する技術分野といった項目に出現するキーワードを上位階層の作成に、課題を解決するための手段」や「発明の実施の形態」といった項目に出現するキーワードを下位階層の作成に利用するというように階層ごとに分類に利用する項目を限定する。分類体系構築補助インタフェース２７００は、こうした文書に含まれる項目とその項目に書かれる文章の専門性をユーザから容易に取り込むことができる。

【０１０５】図２７は文書分類装置２２００が特許の明細書を対象として分類処理を行ったときの分類体系構築補助インタフェース２７００の表示例である。次に、分類指定部２２０１がこの分類体系構築補助インタフェース２７００を介してユーザから補助情報を取り込む処理の流れを図２８の流れ図を用いて説明する。分類指定部２２０１は、ステップ２８０１として画面２７０１にサンプル文書を読み込む。サンプル文書とは文書ＤＢ１０７に格納されている文書群のうちの一つで、図２７では特許の明細書の一例である。続くステップ２８０２ではユーザからサンプル文書中の項目に関する情報を受け取る。ユーザはマウス２７０２を用いて画面２７０１上の文字列を指定することができる。図２７ではマウス２７０２を用いて項目「発明の名称」の文字列をドラッグし、項目として指定した様子を示している。このようにユーザから項目が指定されるとステップ２８０３としてダイアログボックス２７０６を表示し、ユーザの指定した項目が適当かどうかの確認を求める。ステップ２８０４でユーザから確認が取れると、項目とその項目のサンプル文書上の出現位置に関する情報を取り込む（ステップ２８０５）。すなわち、分類指定部２２０１は「発明の名称」という項目と出現位置「１文字目から５文字目まで」という情報を格納する。サンプル文書中の全項目の受取りが終了すると、ステップ２８０６として各項目の出現位置を手掛かりとして項目をサンプル文書での出現順に整列し、これを項目設定画面２７１０の２７１２に表示する（ステップ２７０７）。続くステップ２８０８では、ユーザから全項目の専門性の度合に関する情報を受け取る。分類指定部２２０１は予め用意しておいた専門性を示す数種類の度合をレベルリスト２７１４に表示するので、これを用いてユーザは２７１２から項目の一つを選択し、その内容の一般性、専門性を考慮して２７１４から適切なレベルを一つ選択し、２７１５のOKボタンで確定するという手順で指定が行える。項目設定画面２７１０では「一般的」、「やや一般的」、「どちらともいえない」、「やや専門的」、「専門的」といった項目に関する５種類の専門性の度合が用意されているので、各項目に適切な度合をここから選択する。項目に関する専門性の度合のことを項目レベルと呼び、詳細については後述する。

【０１０６】次に図２９を用いて、分類指定部２２０１が分類体系構築補助インタフェース２７００から取り込まれたユーザの指示に基づいて行う、再分類処理２３０８の詳細を説明する。ステップ２９０１として分類指定部２２０１はユーザから取り込んだ項目とその項目レベルに基づいて階層構築情報を作成する。階層構築情報とは、文書分類部１０３が適切な階層構造の分類結果を構

築するために分類処理２３０３で参照する情報であり、ある階層を構築する場合に分類に利用すべき項目とその項目の重要度を規定するものである。詳細については後述する。

【０１０７】次にステップ２９０２として、この階層構築情報を用いて一時キーワードＤＢを作成する。一時キーワードＤＢとは、ある階層を構築する場合に分類に利用するキーワードとしての重要度を付与したキーワード群の集合を格納したもので、第１の実施例のキーワードＤＢ１０８と同形式である。一時キーワードの作成に関しては後に説明する。これらは分類指定部２２０１がユーザから取り込んだ情報をもとに、適切な階層構造の分類結果を構築するためのデータを準備する処理ステップである。

【０１０８】次に、分類指定部２２０１は文書分類部１０３に対してこれらのデータを用いて適切な階層構造の分類結果を構築するように指示する。文書分類部１０３はステップ２９０３として、第１の実施例におけるキーワードDBの代わりに一時キーワードＤＢを用いて、各キーワードをキーワード群中に含む文書を各キーワードごとにまとめる、単一キーワード処理３０１を行う。

【０１０９】次にステップ２９０４として、単一キーワード分類処理２９０３によって作成された単一キーワードフォルダについて、類似した文書群を含む単一キーワードフォルダの組を統合する関連キーワード処理３０２を行う。第２の実施例では、第１の実施例で説明した図１４のステップ１４０２における統合すべきかどうかを判定するための処理において、適切な階層構造の分類結果を構築するために階層構築情報を用いてワードベクトルの各キーワードに重要度を付与する。この詳細は後述する。

【０１１０】続くステップ３０３では第１の実施例と同様に、関連キーワード処理２９０４によって作成された関連キーワードフォルダについて、類似した文書群を含む関連キーワードフォルダの統合を繰り返す。さらに、作成された単一キーワードフォルダや関連キーワードフォルダ内について、細分類できるかどうかを調べ（第１の実施例のステップ３０４と同様）、細分類可能な場合はフォルダ内を分類する（ステップ２９０５）。ステップ２９０５はフォルダ内に分類されている文書を対象として、図２９におけるステップ２９０２から終了までの処理Bを適用する。ステップ３０６では、すべてのフォルダ内について階層的な分類を繰り返し行う。

【０１１１】次に、階層構築情報（の詳細とその役割について説明する。ここでは、文書分類装置２２００は三階層の分類体系を生成するものとし、項目レベルは「一般的」、「やや一般的」、「どちらでもない」、「やや専門的」、「専門的」の５種類を設定できるようにするものとする。さらに図２８の流れ図に従い、分類体系構築補助インタフェース２７００を介して、図３０の４項

目と各項目レベルがユーザから与えられている。階層構築情報を作成するための規則として、例えば次のようなものを仮定する。

【０１１２】○第一階層構築規則：「一般的」項目レベルの重要度 = a
「やや一般的」項目レベルの重要度 = b
その他の項目レベルの重要度 = ０
○第二階層構築規則：「やや一般的」項目レベルの重要度 = c
「どちらでもない」項目レベルの重要度 = d
「やや専門的」項目レベルの重要度 = e
その他の項目レベルの重要度 = ０
○第三階層構築規則：「やや専門的」項目レベルの重要度 = f
「専門的」項目レベルの重要度 = g
その他の項目レベルの重要度 = ０
例えばこの第一階層構築規則は、第一階層のフォルダを生成する際には「一般的」項目レベルの重要度をa、「やや一般的」項目レベルの重要度をbとし、その他の項目レベルは重要度０、すなわち分類に利用しないことを意味する。例えば、第一階層は一般的な内容に基づき分類するのが好ましいと考えると、「一般的」項目レベルの重要度aを１、「やや一般的」項目レベルの重要度bを0.5、その他の項目レベルの重要度を０とするというようにa～gには０以上の１以下の定数を経験的に決めて、分類指定部２２０１に設定してあるものとする。

【０１１３】上記規則に基づき、図３０の項目レベルから階層構築情報を作成すると以下のようになる。

【０１１４】○第一階層構築情報：(a, 0, 0, b)
○第二階層構築情報：(0, e, 0, c)
○第三階層構築情報：（ ０， ｆ， ｇ， ０ ）
階層構築情報の第一要素は項目「第一章」、第二要素は項目「第二章」、第三要素は項目「第三章」、第四要素は項目「第四章」それぞれの重要度である。すなわち、上例の第一階層構築情報は項目「第一章」を重要度ａ、項目「第四章」を重要度bとして分類に利用し、それ以外の項目は分類には利用しないことを示す。

【０１１５】次に、一時キーワードDBの詳細について説明する。第一階層を構築する場合のステップ２９０２では、第一階層構築情報を用いて次のように一時キーワードDBを作成する。

【０１１６】まず、第一階層構築情報から分類に利用すべき項目は「第一章」と「第四章」であることを読み取る。すなわち、「第二章」と「第三章」は重要度０なので分類に利用せず、それ以外の「第一章」と「第四章」を利用する。

【０１１７】次に、キーワードDB１０８から文書のキーワード群を読み出し、項目「第一章」と項目「第四章」に出現するキーワードだけを取り出して一時キーワード群を作成し、これを一時キーワードDBに登録する。例え

ば、図３１の文書３１００のキーワード群はキーワードDB１０８中のキーワード群３１１０として登録されている。キーワード群３１１０から一時キーワード群を作るには、項目「第一章」３１０１と項目「第四章」３１０３にそれぞれ付随する文章３１０２、３１０４中に出現するキーワードだけを取り出せばよい。すなわち、キーワードa３１０５、キーワードb３１０６、キーワードg３１０７、キーワードh３１０８がこれに相当し、一時キーワード群３１１２を作成し、一時キーワードＤＢ３１１１に登録される。単一キーワード分類２９０３ではこうして作成された一時キーワード群を利用する。

【０１１８】こうした階層構築情報、一時キーワードＤＢを用いて行う関連キーワード分類処理２９０４について説明を加える。第一階層を構築するための関連キーワード処理２９０４では、第一階層構築情報から項目「第一章」に出現するキーワードは重要度a、項目「第四章」に出現するキーワードは重要度bであることを読み取る。次に、各文書の持つキーワードの出現頻度に上記重要度を積算し、第一階層構築時に重視すべきキーワードの重要度を高めて分類処理を行う。このように、ユーザの指定した項目に出現するキーワードを重視して分類することでその階層に適するフォルダが生成しやすくなり、上下関係の適切な分類体系が作り出されることになる。

【０１１９】第２の実施例で説明した分類指定部２２０１について、予めユーザから分類に関する指示を受け取りこれに基づく分類処理を施す、文書分類装置４１００（図４１に示した）に関する第３の実施例を説明する。

【０１２０】ユーザがキーボード１１０、マウス１１１といった入力装置を用いて分類したい文書群を指定し、分類処理の実行を指示すると、文書分類装置４１００が起動され、図３２の流れ図に基づく処理が施される。まずステップ２３０１としてデータ管理部１０１が文書ＤＢ１０７にユーザの指定した文書群を格納する。

【０１２１】続くステップ２３０２では、単語検出部１０２が文書群から単語キーワードを検出し、単語辞書１０５に格納する。こうして分類処理を行うためのデータが用意できると、ステップ２３０５として分類指定部２２０１に対するユーザからの指示を確認し、指示がない場合には待ち続ける。ユーザから指示があると文書分類部１０３にユーザの指示を受け渡して分類処理３２０１を実行させる。

【０１２２】分類処理３２０１の詳細は後述する。次にステップ２３０４として分類結果出力部１０４が分類結果を表示する。さらに分類指定部２２０１は分類結果を参照したユーザからの指示を確認し（ステップ２３０５）、指示がない場合には処理を終了する。指示がある場合にはステップ２３０６で指示内容を解釈し、フォルダ数指定インタフェース２５００への指示の場合はステップ２３０７の再分類処理を施し、分類体系構築補助インタフェース２７００への指示の場合はステップ２３０８の再分類処理を適用する。ステップ２３０５で再びユーザからの分類に関する指示がある例としては、予めユーザが指定した分類結果が予想に反して意向に沿わなかった場合などが考えられる。

【０１２３】図３３に分類処理３２０１の具体的な処理の流れを示す。まずステップ３３０１として、フォルダ数指定インタフェース２５００に対する指示か、分類体系構築補助インタフェース２７００に対する指示かを調べ、後者の場合は分類体系構築補助インタフェース２７００の再分類処理２３０８を実行して処理を終了する。前者の場合はステップ２３０３で文書分類部１０３が図３の流れ図に従って分類体系を生成し文書を分類する。次にステップ３３０２として、分類結果がユーザの指示条件を満たすかどうかを確認する。フォルダ数指定インタフェース２５００へのユーザからの指示はフォルダ数あるいは平均文書数が指定されるものなので、これに見合った分類結果が得られているかどうかを調べ、得られた場合は処理を終了する。ここでユーザからの指示が満たされない場合は、フォルダ数指定インタフェース２５００の再分類処理２３０７を実行してユーザの指示に見合うように再び分類処理２３０３を実行する。

【０１２４】図１に示した第１の実施例の文書分類装置１００における分類結果出力部１０４に、ユーザが分類結果をブラウジングする手段を設けた文書分類装置に関する第４の実施例を説明する。

【０１２５】第１の実施例の文書分類装置は予め設定した分類体系に基づいて分類する場合と異なり、分類結果として生成されるフォルダが全く未知である。そこでユーザが分類結果を参照して所望の文書を検索する場合には、まずどのようなフォルダがあるのかを調べ、その上で所望の文書が分類されていそうなフォルダを選びフォルダ内をブラウジングする。しかし、所望の文書に辿り着くまでの手間が大きいと分類体系が固定化されている方がブラウジングしやすいということになる。これでは分類対象の文書に応じて適切な分類体系を生成する本発明が有効に活用されない。そこで図３４のように、第１の実施例の文書分類装置に検索支援部３４０１を設けた文書分類装置３４００の構成を取り、分類結果のブラウジングの支援が行えるようにする。検索支援部３４０１はフォルダ検索支援機能３４０２と文書検索支援機能３４０３という二つの処理機能から構成される。

【０１２６】ユーザがキーボード１１０、マウス１１１といった入力装置を用いて分類したい文書群を指定し、分類処理の実行を指示すると、文書分類装置３４００が起動され、図３５の流れ図に基づく処理が施される。まずステップ２３０１としてデータ管理部１０１が文書ＤＢ１０７にユーザの指定した文書群を格納する。

【０１２７】続くステップ２３０２では、単語検出部１０２が文書群から単語キーワードを検出し、単語辞書１

０５に格納する。こうして分類処理を行うためのデータが用意できると、ステップ２３０３で文書分類部１０３が図３の流れ図に従って分類体系を生成し、文書を分類する。分類結果出力部１０４はこの分類結果を図３６のようなインタフェース画面３６００に表示して、ユーザに提示する（ステップ２３０４）。

【０１２８】本図では３６０１に第一階層のフォルダ名が、３６０２に第二階層のフォルダ名が、３６０３に第３階層のフォルダ名が表示されている。ここまでは第１および第２の実施例と同一の処理ステップである。続くステップ３５０１では、分類結果出力部１０４が表示した分類状況のうち、任意のフォルダあるいは文書をユーザが選択したかを調べる。もし何も選択がないまま、ユーザが終了ボタン３６０４で分類結果の終了を指示した場合は処理を終える（ステップ３５０２）。

【０１２９】選択した場合は、ステップ３５０３で分類結果出力部１０４がフォルダを選択しているのか文書を選択しているのかを調べ、フォルダの場合は検索支援部３４０１のフォルダ検索支援機能３４０２にフォルダ検索支援処理３５０４を指示し、文書の場合は検索支援部３４０１の文書検索支援機能３４０３に文書検索支援処理３５０５を指示する。フォルダ検索支援処理３５０４および文書検索支援処理３４０５はブラウジングしやすいように分類結果の加工を行うもので、詳細については後述する。さらに、ユーザが検索支援ボタン３６０５を押し、分類結果のブラウジングの支援を要求した場合には、検索支援部３４０１が分類結果出力部１０４に加工した分類結果を受け渡し、新たな分類結果が表示される。

【０１３０】次に、検索支援部３４０１の提供する、フォルダ検索支援機能３４０２および文書検索支援機能３４０３の詳細について述べる。

【０１３１】フォルダ検索支援機能３４０２とは、ユーザが参照しようとして選択したフォルダとの類似度に基づいてその他のフォルダを評価し、この類似度順にフォルダの並び換えを行うものである。図３６は分類結果出力部１０４が出力する分類結果のインタフェース画面３６００と、検索支援部３４０１が出力する補助情報画面３６１０を示している。インタフェース画面３６００には、第一階層３６０１、第二階層３６０２、第三階層３６０３のフォルダ名が表示され、ユーザが選択し参照できるようになっている。ここでは説明を簡略化するため、フォルダ名を抽象化して「フォルダa」などのように示しているが、実際には１個以上のキーワードによって構成される。補助情報画面３６１０には、フォルダ数３６１１や平均文書数３６１２といった分類結果に関する補助情報が表示される。ユーザがまだ何も選択していない状態のときは、画面３６１３には第一階層に生成されたフォルダがその文書数に応じた大きさの円で表現され、文書数の多い順番に整列される。ユーザはこれらの

情報を手掛かりとして所望のフォルダを選択し、フォルダに分類された文書を参照する。

【０１３２】例えば、ユーザが第一階層３６０１の中からフォルダaを選択すると、分類結果出力部１０４はこれを検索支援部３４０１のフォルダ検索支援機能３４０２に知らせ、フォルダ検索支援処理３５０４を実行する。フォルダ検索支援処理３５０４では、まずフォルダaと第一階層のその他のフォルダとの類似度を調べる。フォルダ間の類似度は、「フォルダaに分類された文書と一致する文書が多いフォルダほど類似度が高い」と判断する。これは第１の実施例で説明した文書分類部１０３の関連キーワード処理３０２の中で求められる。すなわち、第４の実施例では図３５のステップ３５０１の分類処理がこれに相当する。

【０１３３】フォルダ検索支援処理３５０４ではこのフォルダ間の類似度が必要になるので、第４の実施例の分類処理３５０１ではフォルダ間で一致する文書数に関する情報を保存しておき、フォルダ検索支援機能３４０２に情報を提供する。具体的には図９のステップ９０２でフォルダ間一致文書数管理テーブル１００１の内容の保存を行う。

【０１３４】例えば、フォルダaとその他のフォルダとの類似度を調べるには、まずフォルダ間一致文書数管理テーブル１００１を参照して、フォルダ識別番号１００２、１００３からフォルダaに対応する識別番号のカラムだけを取り出す。フォルダ間一致文書数管理テーブル１００１は一致文書数を手掛かりとして降順に整列されているので、取り出したカラムのうち、テーブルの上位に位置するカラムほどフォルダaとの一致文書数が多く、類似度が高いことになる。すなわち、フォルダaとの類似度順に第一階層のフォルダに関する情報を取り出すことができる。フォルダ検索支援機能３４０２はこれらの情報を用いて、補助情報画面３６１０の画面３６１３を図３７の補助情報画面３７００の画面３７０１のように書き換える。すなわち、フォルダaと類似する順番に第一階層のフォルダを並び換え、ユーザがフォルダaと類似したフォルダに関する情報を得やすくなるようにしている。

【０１３５】次に、文書検索支援機能３４０３について説明する。

【０１３６】文書検索支援機能３４０３とは、フォルダに分類された文書のうち、ユーザが不適切と判断した文書と類似する文書を調べ、これをフォルダから除外することで所望の文書を検索する操作を支援するものである。図３８は分類結果出力部１０４が出力する分類結果のインタフェース画面３８００と、検索支援部３４０１が出力する補助情報画面３８１０を示している。インタフェース画面３８００には、第一階層３８０１にフォルダ名、３８０２にはユーザによって選択されたフォルダaに関する第二階層のフォルダ名、３８０３にはサブフ

ォルダaに関する文書名一覧が表示され、ユーザが３８０３から文書aを選択し、文書の内容を表示する画面３８０４で文書aを参照している。

【０１３７】補助情報画面３８１０の３８１１には、現在ユーザが参照中の文書aをはじめとするサブフォルダa中の文書と、各文書の内容の先頭部分が一覧表示されている。ユーザはこれらの情報を手掛かりとして分類結果をブラウジングする。情報を参照した結果、例えば「文書aは不要」とユーザが判断し、不要な文書を指定する消去ボタン３８１２を押すと、検索支援部３４０１はこの情報を文書検索支援機能３４０３に知らせる。文書検索支援機能３４０３は、図３９の流れ図に従って文書検索支援処理３５０５を行う。不要と判断された文書aと類似する文書を調べるために、第１の実施例で述べたワードベクトル間の距離計算処理１４０５を用いる。そのための準備として、フォルダ内文書の距離計算処理１３００をサブフォルダaについて行う。

【０１３８】次に、図１４のステップ１４０１でワードベクトルテーブル１５００の初期化を行い、各文書についてワードベクトルの作成を繰り返す（ステップ１４０２、１４０３）。ここまでは第１の実施例と同一の処理である。続くステップ３９０１では、文書aとその他の各文書とについて第１の実施例のワードベクトル間の距離計算処理１４０５を行う。文書間の距離は類似度であることから、文書aと各文書との類似度が求められる。さらにステップ３９０２として、文書aとの距離がしきい値Tα以下の文書、すなわち文書aとの類似度が近いものを調べ、これらを不適切な文書の候補とみなす（ステップ３９０３）。全文書についてこの処理を繰り返し（ステップ３９０４）、文書aと類似度の高いすべての文書を選び出す。

【０１３９】検索支援部３４０１は、不適切な文書の候補と判断された文書に関する情報を分類結果出力部１０４に受け渡す。この後、ユーザが検索支援ボタン３８０５を指定すると、文書aと類似した文書をフォルダから除外して、新たに図４０の分類結果のインタフェース画面４０００を作成する。サブフォルダaは文書aおよび文書aと類似すると判断された文書が除かれて文書数が図３８の９６件（３８０６）から図４０の７１件（４００１）になり、文書一覧３８０３は文書aおよび文書aと類似すると判断された文書d、文書fなどの２５件が除かれて４００２のように変更されている。

【０１４０】

【発明の効果】本発明によれば、以下の効果が得られる。

【０１４１】（１）既定の分類体系に依存することなく、文書を分類することができる。

【０１４２】（２）階層的な分類体系を自動的に生成することができる。

【０１４３】（３）分類結果群の代表名称を付与することができる。代表名称は分類結果群との関連度の高いものから順番に付与する。

【０１４４】その結果、ユーザが特に興味の対象を指定しなくても、大量の文書の中から所望の文書を見つけやすくすることができる。あるいは興味の分野は定まっていても、それを代表するキーワードが思いつかない場合の助けとなる。

【図面の簡単な説明】

【図１】本発明の一実施例のシステム構成図である。

【図２】本実施例のキーワード説明図である。

【図３】本実施例の文書分類処理の流れ図である。

【図４】本実施例の単一キーワード分類処理の流れ図である。

【図５】本実施例のキーワードテーブルのデータ構造を表わす図である。

【図６】本実施例のキーワードテーブルの具体例を示す図である。

【図７】本実施例のフォルダテーブルのデータ構造を表わす図である。

【図８】本実施例の単一キーワード分類処理の具体例を示す図である。

【図９】本実施例の関連キーワード分類処理の流れ図である。

【図１０】本実施例のフォルダ間一致文書数管理テーブルのデータ構造を表わす図である。

【図１１】本実施例のフォルダ間一致文書数管理テーブルの具体例を示す図である。

【図１２】本実施例のフォルダ統合時に生じるフォルダ間一致文書数管理テーブルの更新処理の具体例を示す図である。

【図１３】本実施例のフォルダ内文書情報の距離計算の流れ図（その１）である。

【図１４】本実施例のフォルダ内文書情報の距離計算の流れ図（その２）である。

【図１５】本実施例のワードベクトルテーブルの具体例を示す図である。

【図１６】本実施例の関連キーワード分類処理結果の具体例を示す図である。

【図１７】本実施例の細分類処理の流れ図である。

【図１８】本実施例の細分類処理結果の具体例を示す図である。

【図１９】本実施例の分類階層管理テーブルのデータ構造を表わす図である。

【図２０】本実施例の最終的な分類結果例を示す図である。

【図２１】本実施例の分類結果の画面表示の具体例を示す図である。

【図２２】本発明の第２の実施例のシステム構成図である。

【図２３】第２の実施例の文書分類装置の流れ図であ

る。

【図２４】第２の実施例の分類結果の画面表示の具体例を示す図である。

【図２５】第２の実施例のフォルダ数指定インタフェースの具体例を示す図である。

【図２６】第２の実施例のフォルダ数指定インタフェースにおける再分類処理の流れ図である。

【図２７】第２の実施例の分類体系構築補助インタフェースの具体例を示す図である。

【図２８】第２の実施例の分類体系構築補助インタフェースにおいて項目設定を行う画面の具体例を示す図である。

【図２９】第２の実施例の分類体系構築補助インタフェースにおける再分類処理の流れ図である。

【図３０】第２の実施例の項目と項目レベルの具体例を示す図である。

【図３１】第２の実施例の一時キーワードDBの説明図である。

【図３２】第２の実施例の分類指定部２２０１を図２３とは別に適用した場合の、文書分類装置の流れ図である。

【図３３】第２の実施例の図３２における分類処理の詳

細を示す図である。

【図３４】第３の実施例のシステム構成図である。

【図３５】第３の実施例の文書分類装置の流れ図である。

【図３６】第３の実施例の分類結果の画面表示の具体例およびフォルダ検索支援機能によって提示される補助情報画面の具体例を示す図である。

【図３７】第３の実施例の補助情報画面がユーザからの指示に基づいて変化した具体例を示す図である。

【図３８】第３の実施例の分類結果の画面表示の具体例および文書検索支援機能によって提示される補助情報画面の具体例を示す図である。

【図３９】第３の実施例の文書検索支援機能による検索支援処理の流れ図である。

【図４０】第３の実施例の文書検索支援機能によって分類結果が変化した具体例を示す図である。

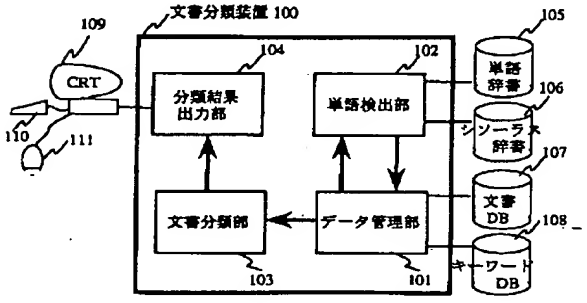
【図４１】第３の実施例のシステム構成図である。

【符号の説明】

１００…文書分類装置、１０２…単語検出部、１０３…文書分類部、３０１…単一キーワード分類処理、３０２…関連キーワード分類処理、１３００…フォルダ内文書の距離計算

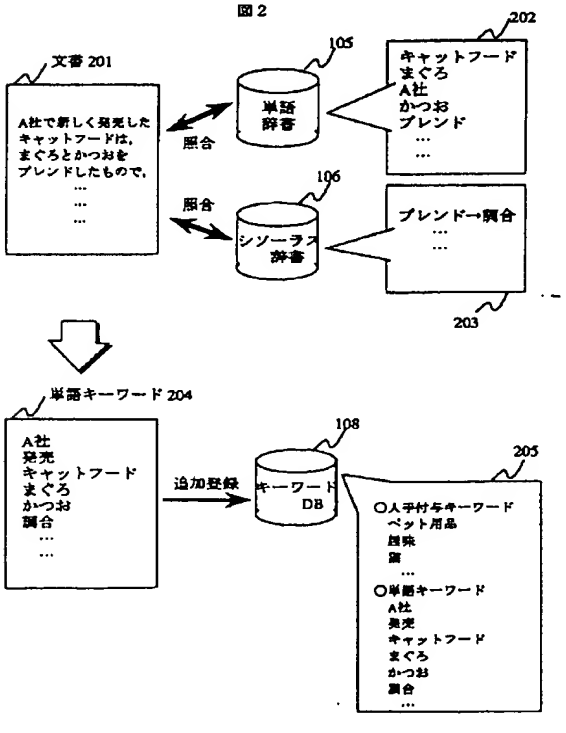
【図１】

図１



【図２】

図２



【図１０】

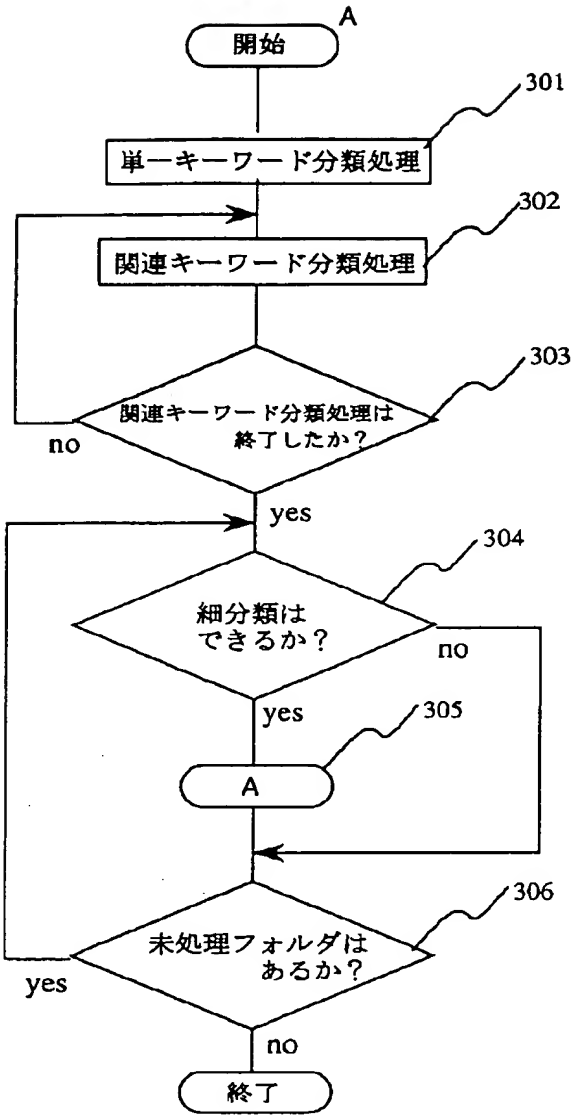
図１０

フォルダ間一致文書数管理テーブル 1001

1002 フォルダ識別番号	1003 フォルダ識別番号	1004 一致文書数	1005 統合済フラグ
0	1	6	済

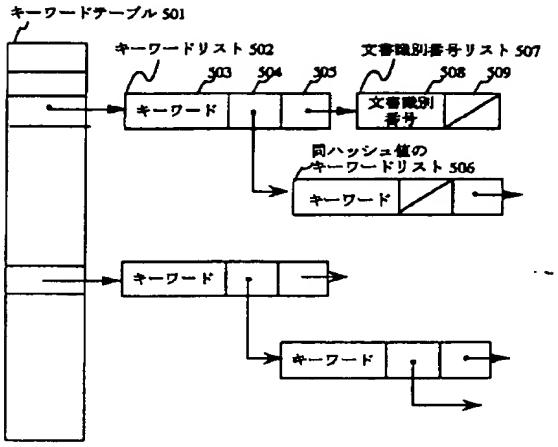
【図３】

図３



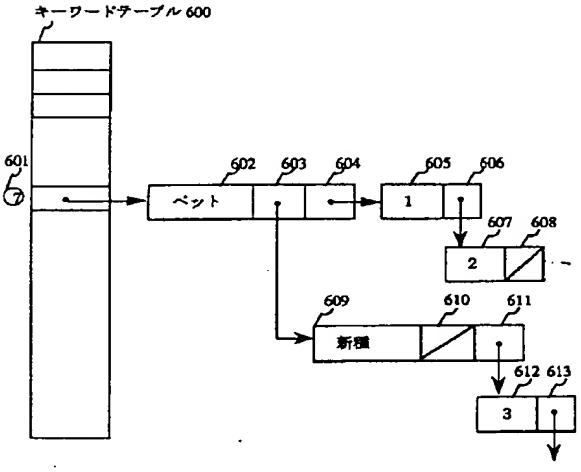
【図５】

図５



【図６】

図６



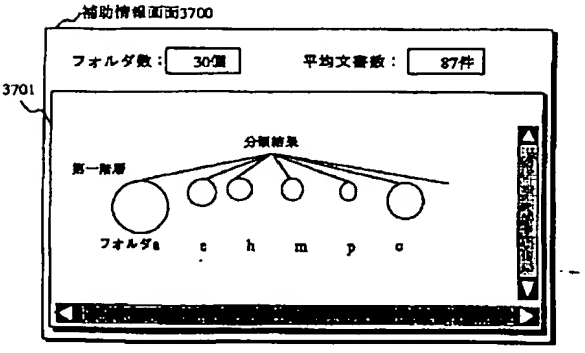
【図３０】

図３０

項目	項目レベル
第一章	一般的
第二章	やや専門的
第三章	専門的
第四章	やや一般的

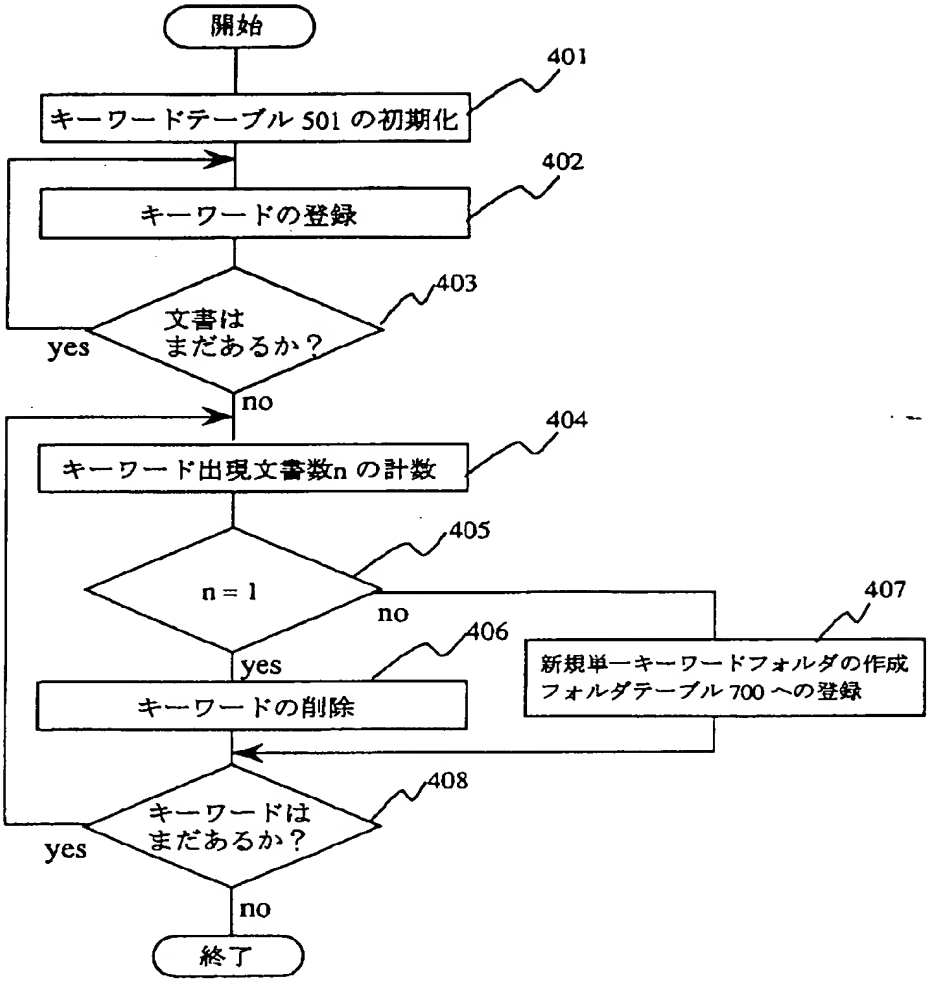
【図３７】

図３７



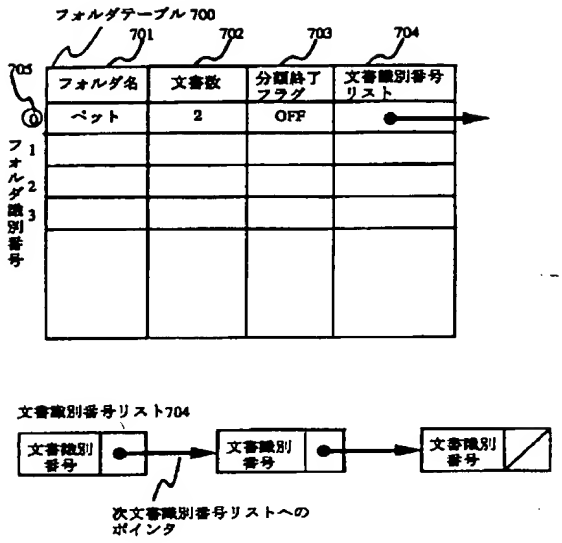
【図４】

図４


開始
401
キーワードテーブル 501 の初期化
402
キーワードの登録
403
文書は
まだあるか？
yes
no
404
キーワード出現文書数n の計数
405
n = 1
no
yes
407
新規単一キーワードフォルダの作成
フォルダテーブル 700 への登録
406
キーワードの削除
408
キーワードは
まだあるか？
yes
no
終了

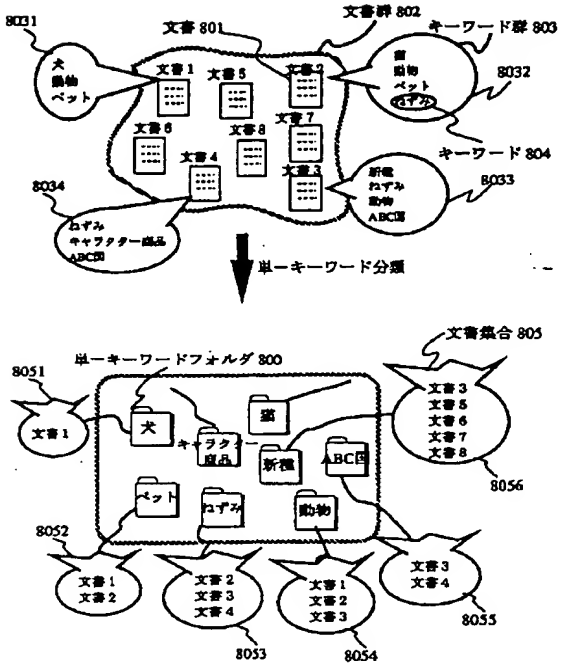
【図７】

図７



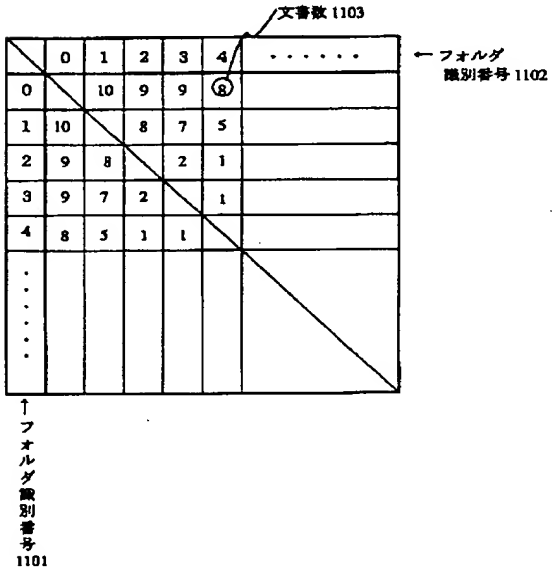
【図８】

図８



【図１１】

図１１



【図１２】

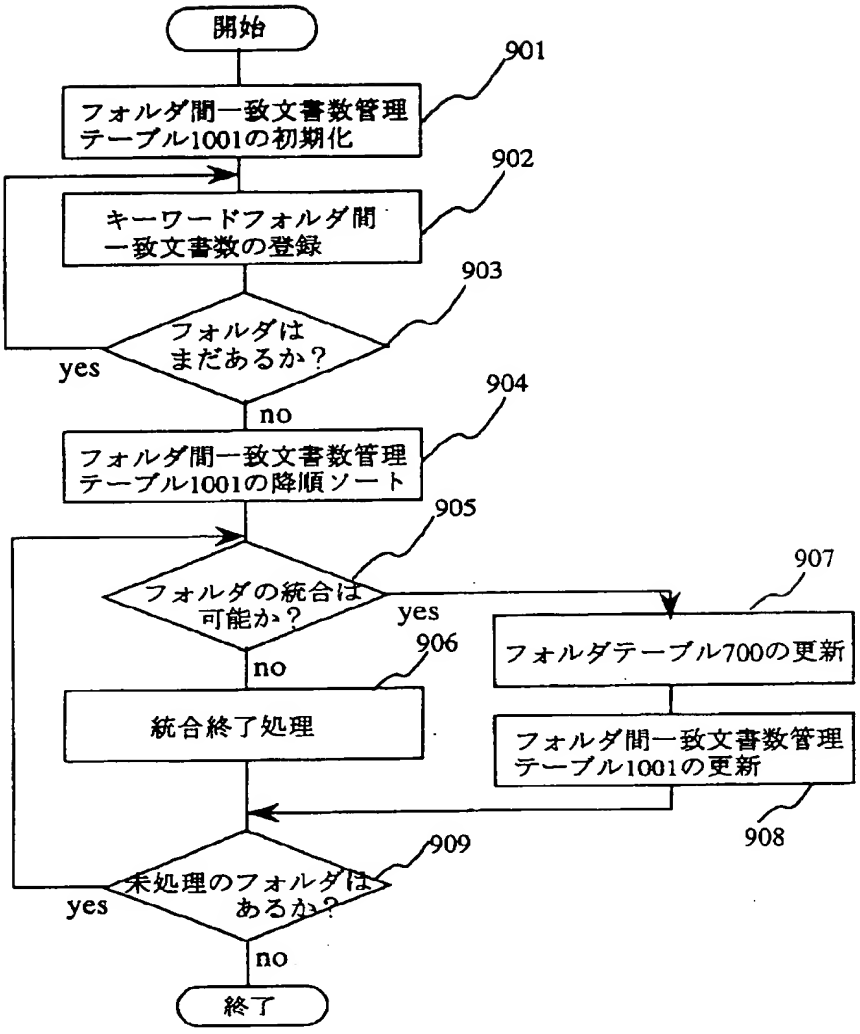
図１２

フォルダ間一致文書数管理テーブル 1200

	フォルダ識別番号 1201	フォルダ識別番号 1202	一致文書数 1203	統合済フラグ 1204	
0	0	1	10	済	1205
1	3	5	9		1214
2	5	1	9	不可	1206
3	0	2	8	不可	1207
4	0	3	8	不可	1208
5	2	5	8		
6	5	3	7		1209
7	3	1	5	不可	1210
8	0	4	2	不可	1211
9	1	2	2	不可	
10	6	5	1		
11	2	4	1		1212
12	1	4	1	不可	1213
13	0	5	1	不可	

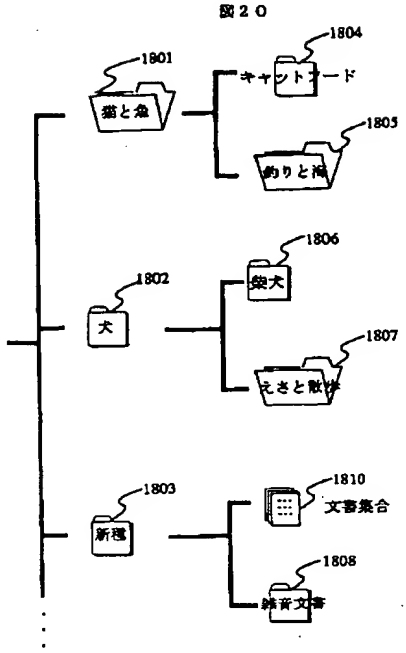
【図９】

図９



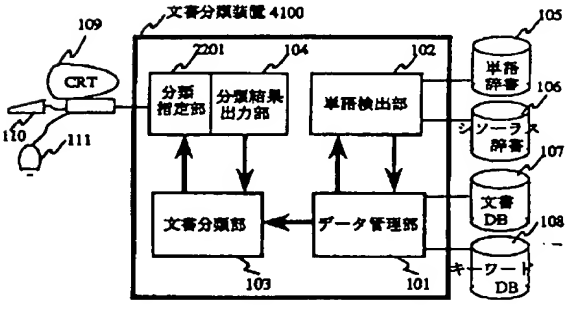
【図２０】

図２０



【図４１】

図４１



【図１３】

図１３

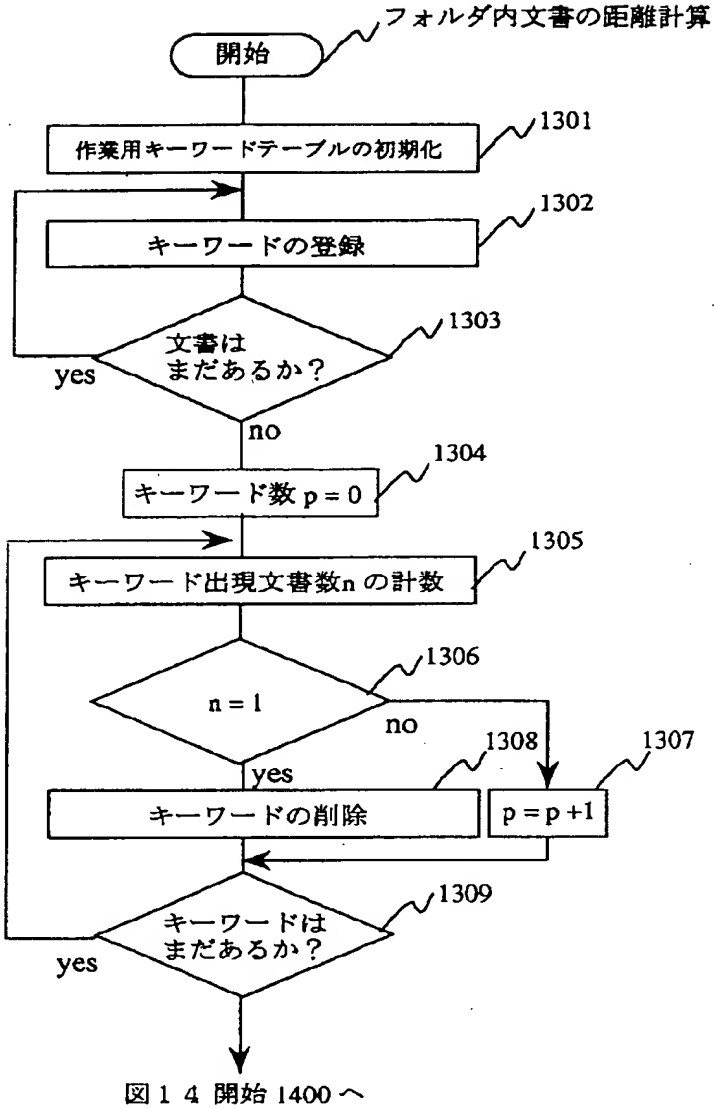
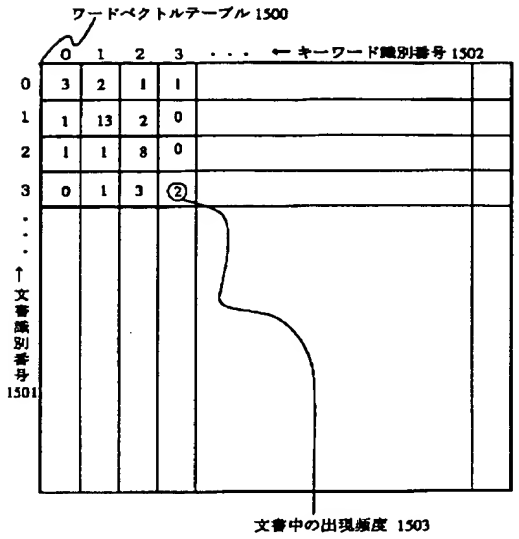


図１４ 開始 1400 へ

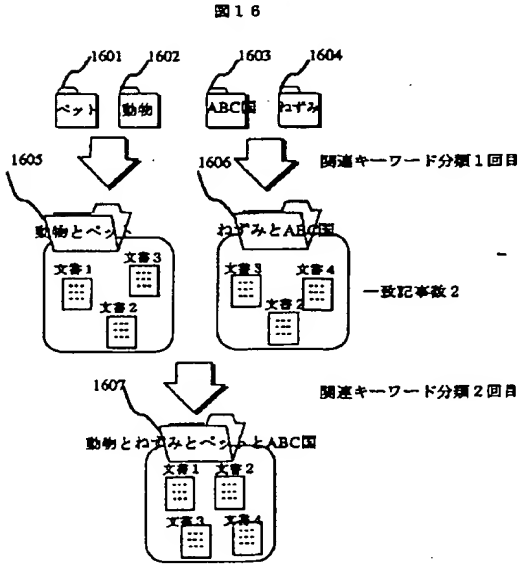
【図１５】

図１５



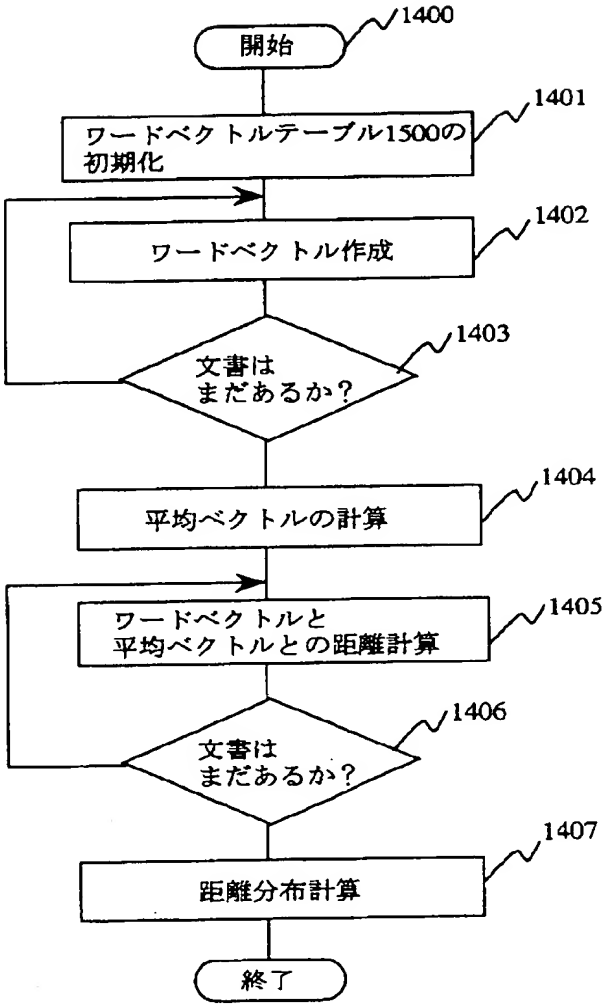
【図１６】

図１６



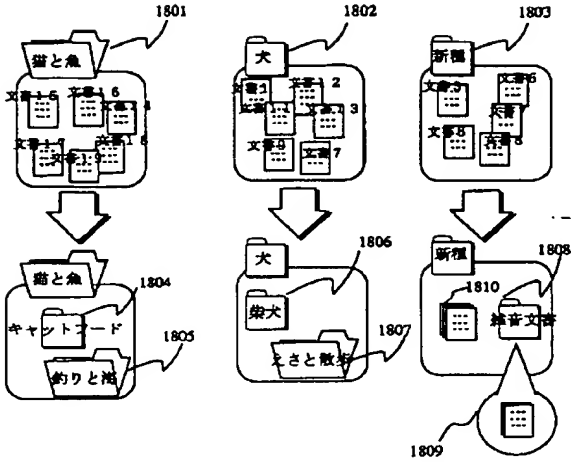
【図１４】

図１４



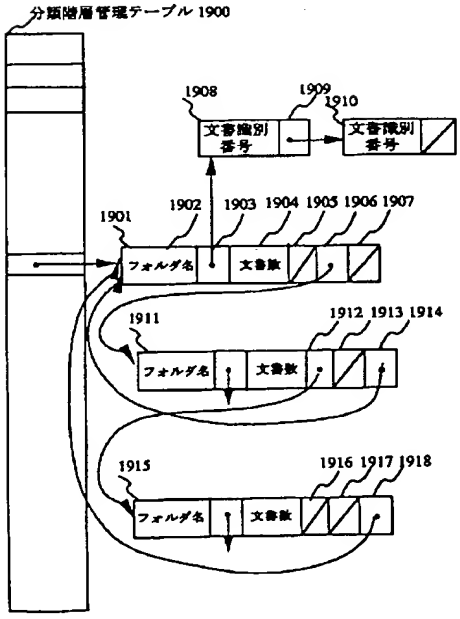
【図１８】

図１８



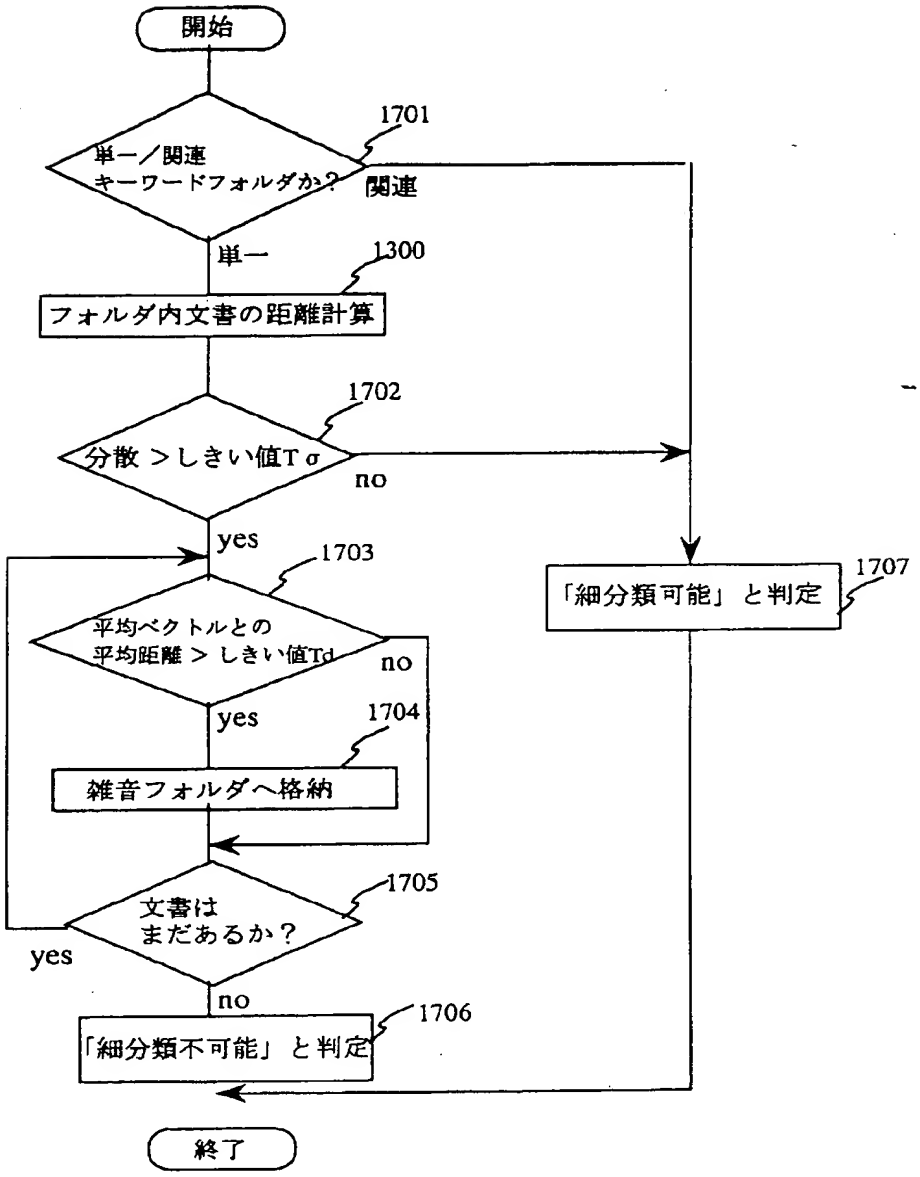
【図１９】

図１９



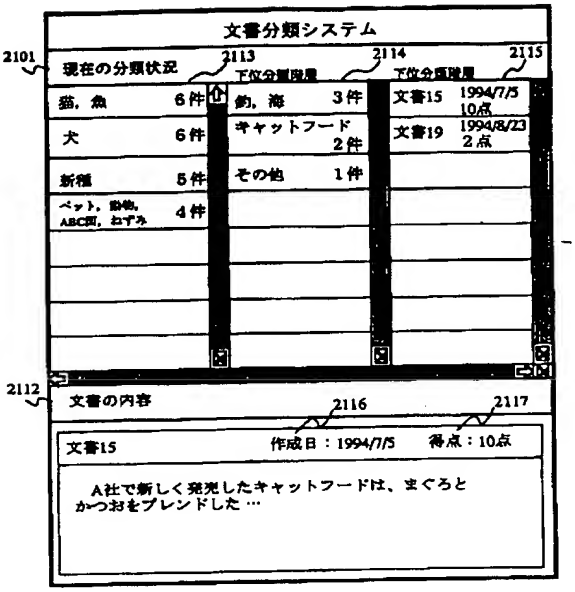
【図１７】

図１７


開始
1701
単一／関連
キーワードフォルダか？
関連
単一
1300
フォルダ内文書の距離計算
1702
分散 ＞しきい値Tσ
no
yes
1703
平均ベクトルとの
平均距離 ＞ しきい値Td
no
yes
1704
雑音フォルダへ格納
1705
文書は
まだあるか？
yes
no
1706
「細分類不可能」と判定
1707
「細分類可能」と判定
終了

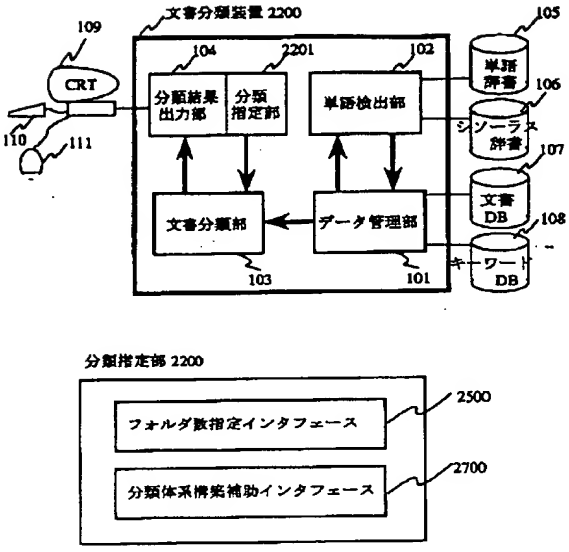
【図２１】

図２１



【図２２】

図２２



【図２７】

【図２５】

図２５

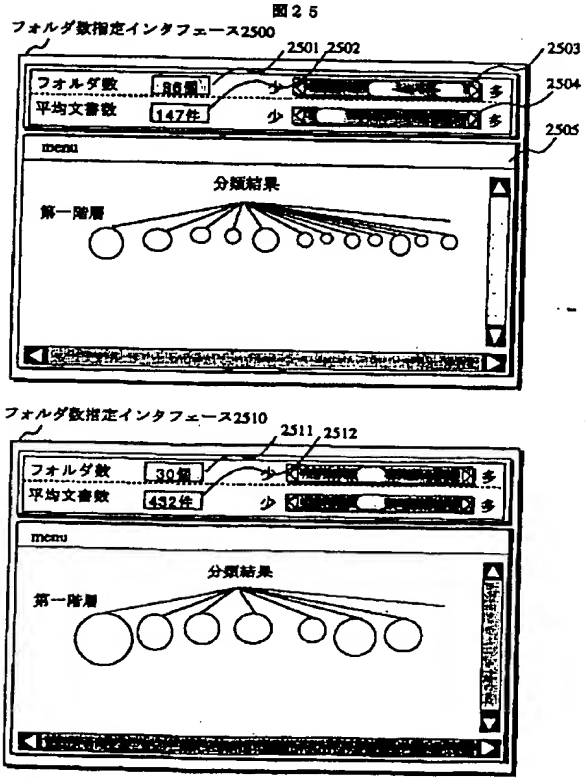
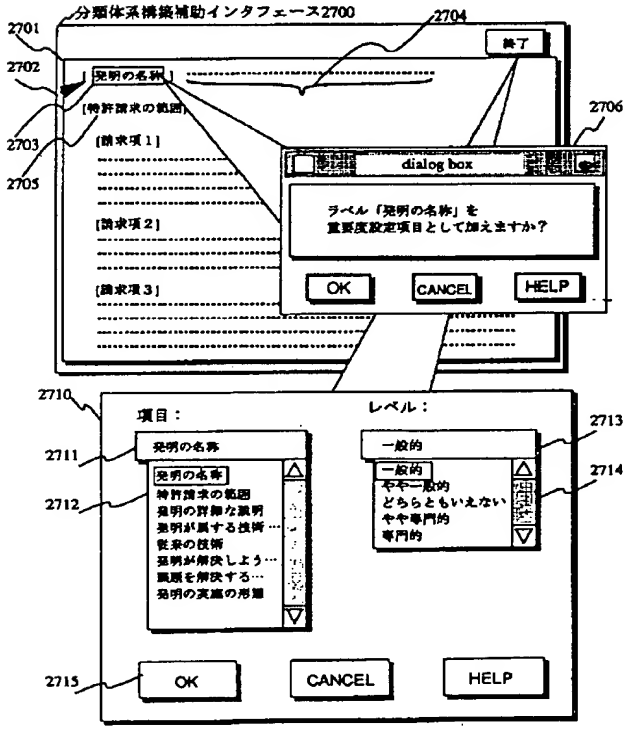
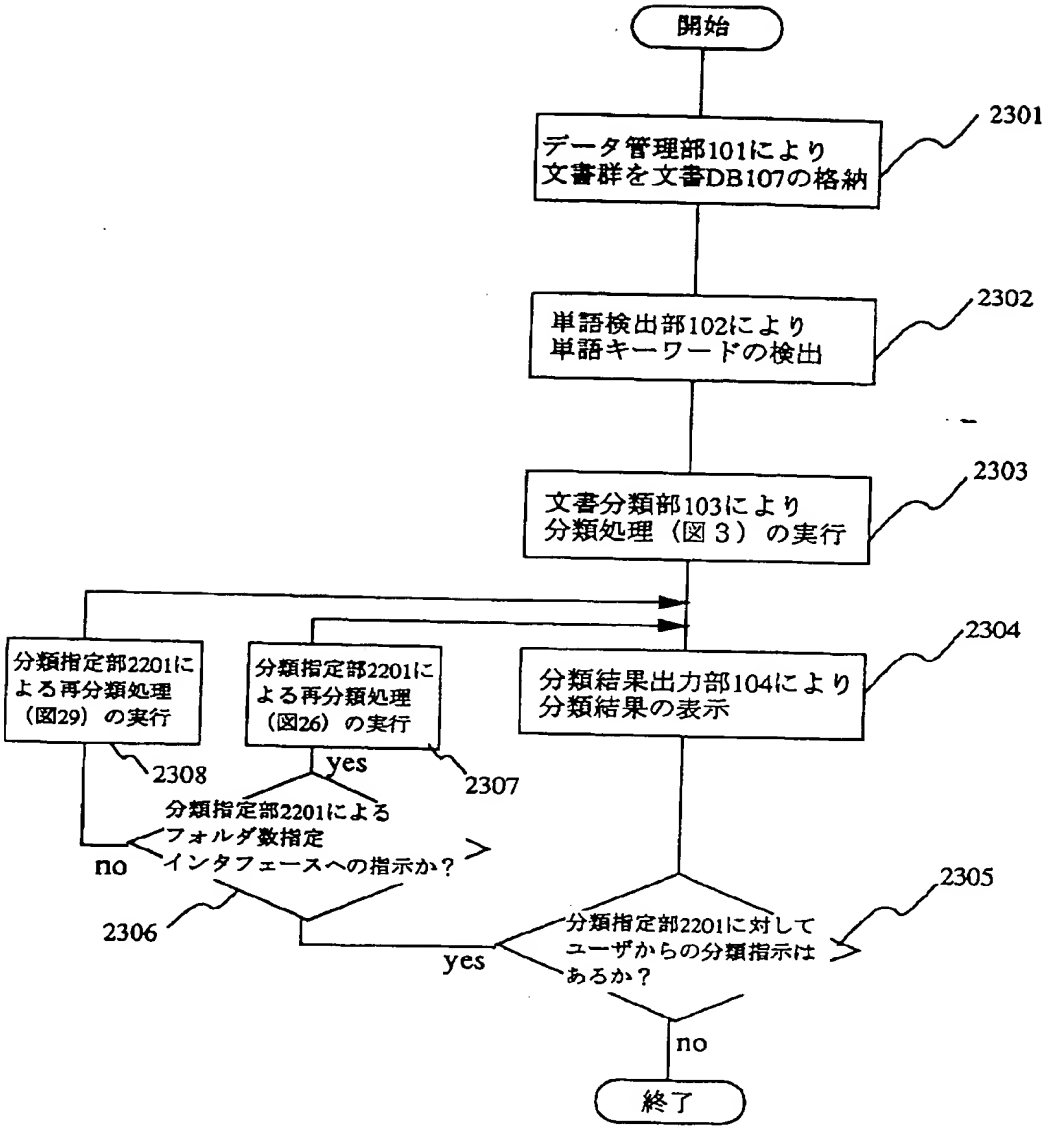


図２７



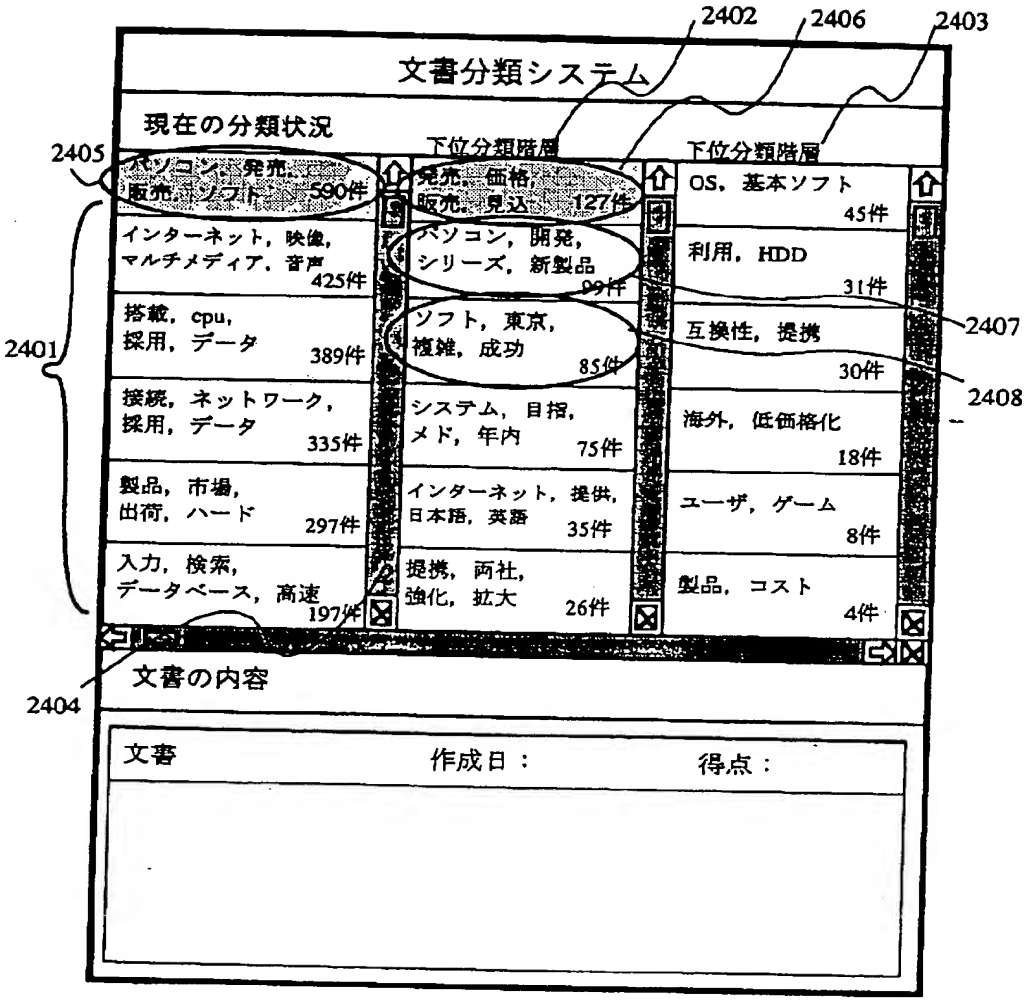
【図２３】

図２３


開始
データ管理部101により
文書群を文書DB107の格納
2301
単語検出部102により
単語キーワードの検出
2302
文書分類部103により
分類処理（図３）の実行
2303
分類指定部2201による再分類処理（図29）の実行
分類指定部2201による再分類処理（図26）の実行
分類結果出力部104により
分類結果の表示
2304
2308
yes
2307
分類指定部2201による
フォルダ数指定
インタフェースへの指示か？
no
2306
yes
分類指定部2201に対して
ユーザからの分類指示は
あるか？
2305
no
終了

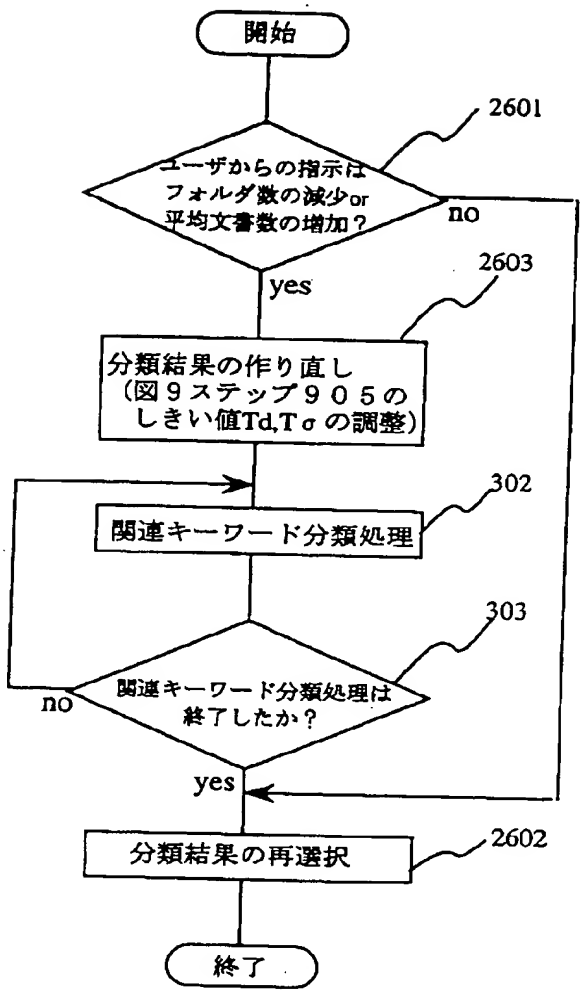
【図２４】

図２４


2402
2406
2403
文書分類システム
現在の分類状況
下位分類階層
下位分類階層
2405
パソコン，発売，
販売，ソフト　590件
発売，価格，
販売，見込　127件
OS，基本ソフト
45件
インターネット，映像，
マルチメディア，音声
425件
パソコン，開発，
シリーズ，新製品
99件
利用，HDD
31件
2407
搭載，cpu，
採用，データ
389件
ソフト，東京，
複雑，成功
85件
互換性，提携
30件
2401
2408
接続，ネットワーク，
採用，データ
335件
システム，目指，
メド，年内
75件
海外，低価格化
18件
製品，市場，
出荷，ハード
297件
インターネット，提供，
日本語，英語
35件
ユーザ，ゲーム
8件
入力，検索，
データベース，高速
197件
提携，両社，
強化，拡大
26件
製品，コスト
4件
2404
文書の内容
文書
作成日：
得点：

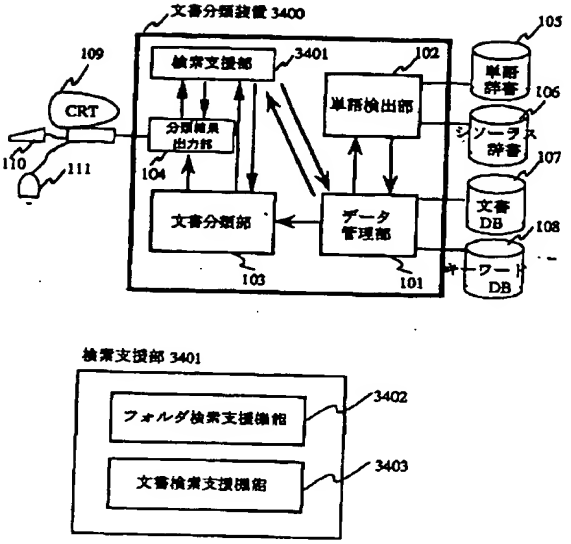
【図２６】

図２６



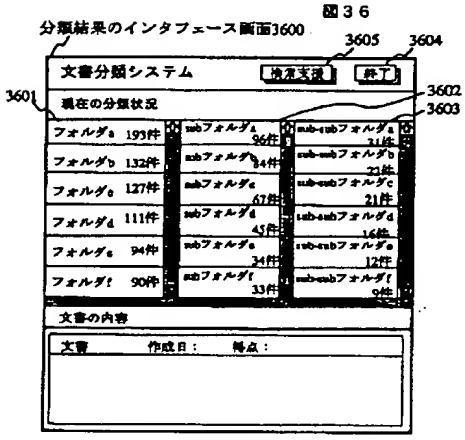
【図３４】

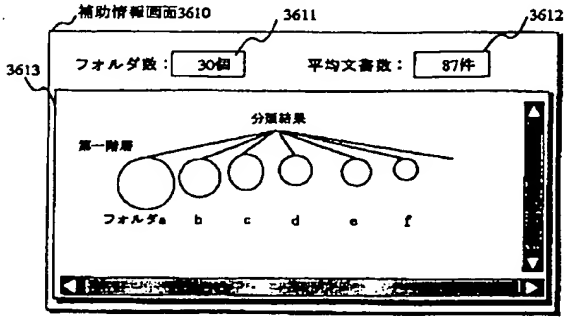
図３４



【図３６】

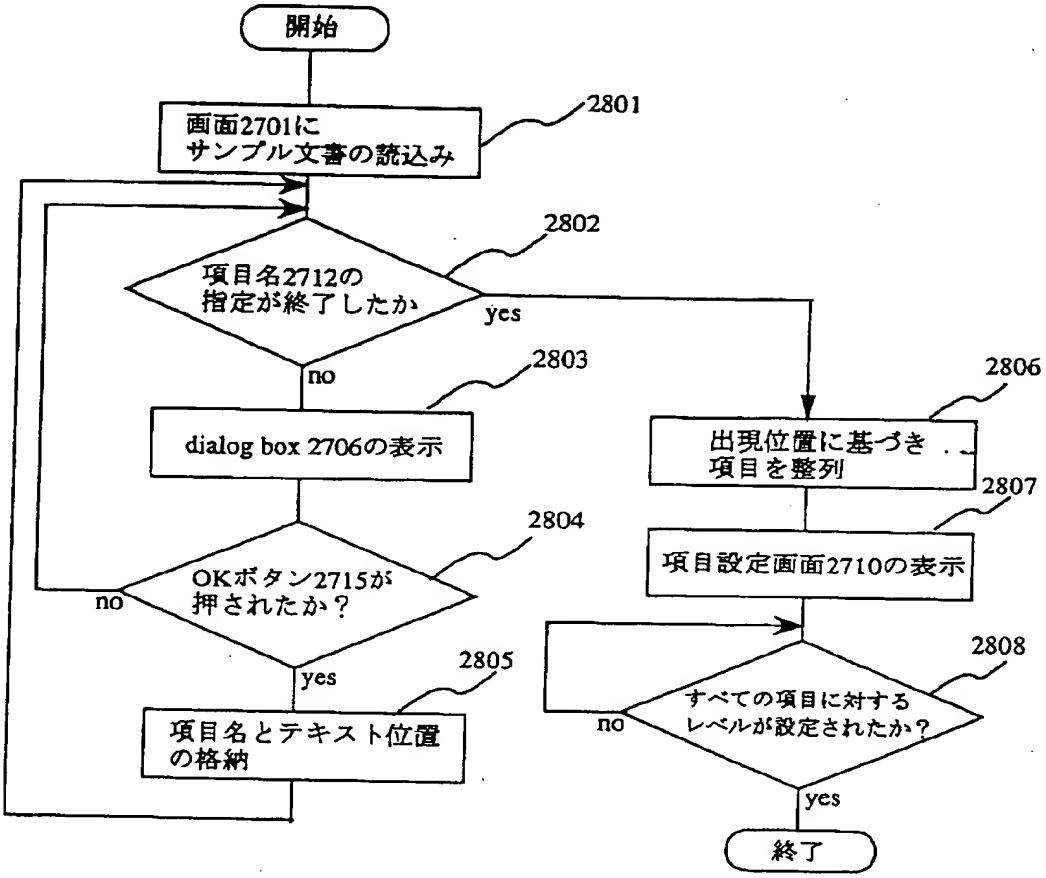
図３６





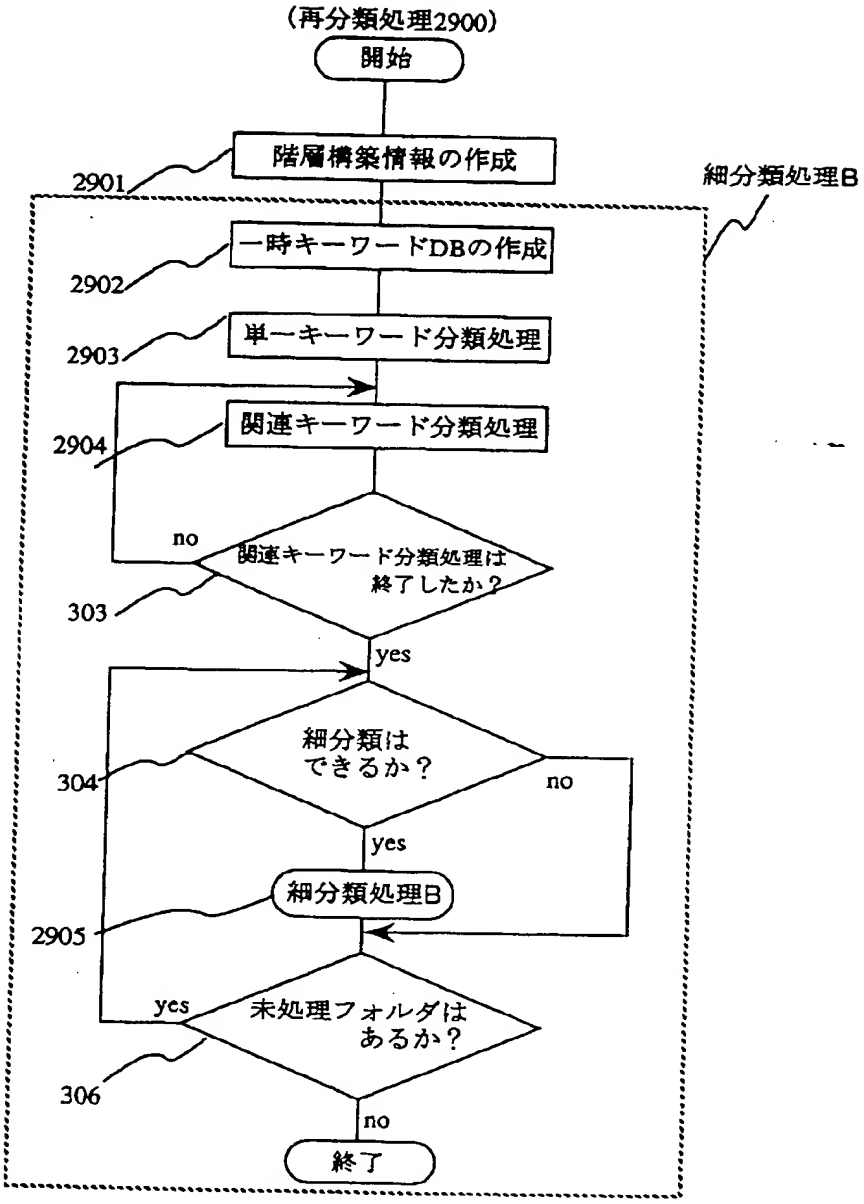
【図２８】

図２８


開始
画面2701に
サンプル文書の読込み
2801
項目名2712の
指定が終了したか
2802
yes
no
2803
dialog box 2706の表示
2804
OKボタン2715が
押されたか？
no
yes
2805
項目名とテキスト位置
の格納
2806
出現位置に基づき
項目を整列
2807
項目設定画面2710の表示
2808
すべての項目に対する
レベルが設定されたか？
no
yes
終了

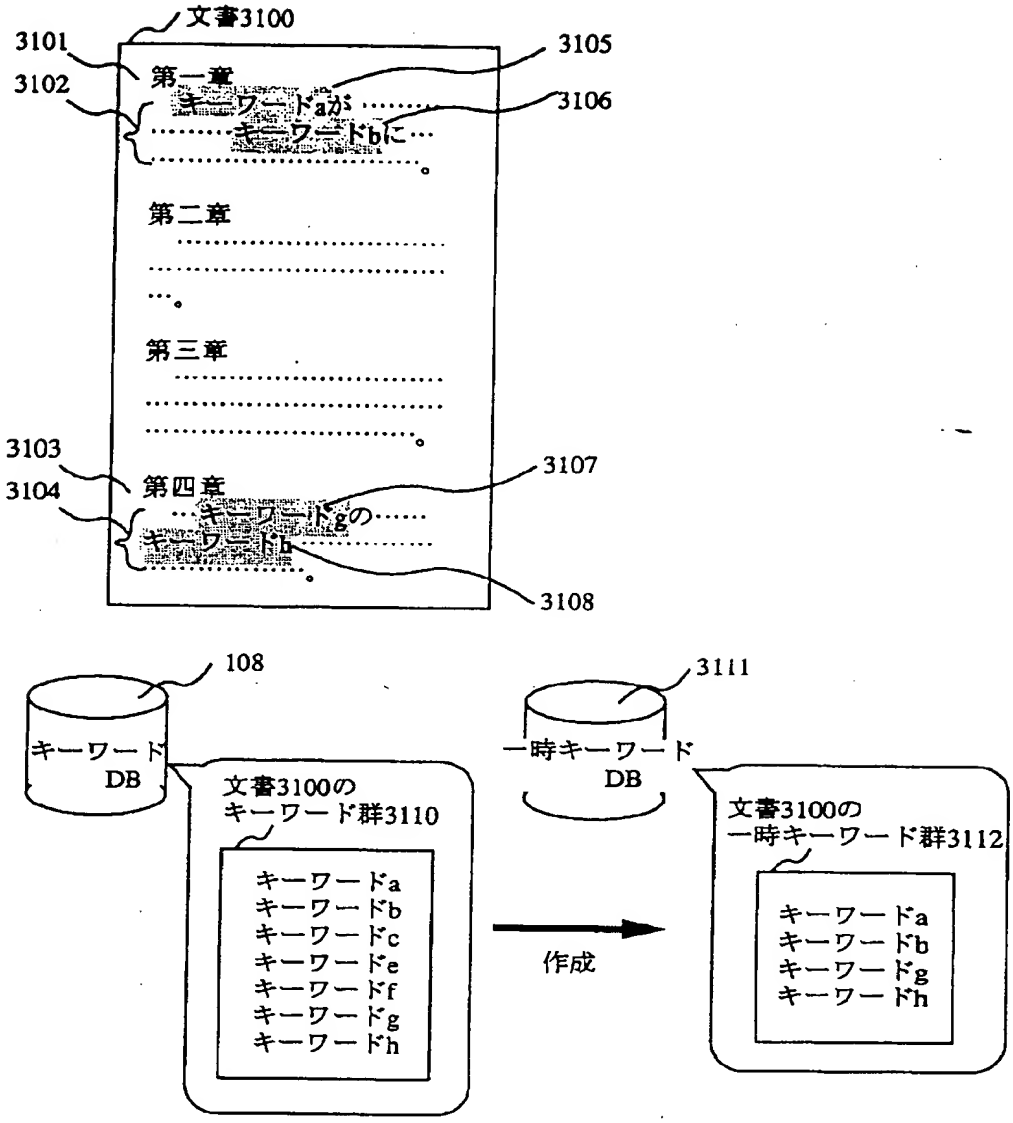
【図２９】

図２９


(再分類処理2900)
開始
階層構築情報の作成
2901
細分類処理B
一時キーワードDBの作成
2902
単一キーワード分類処理
2903
関連キーワード分類処理
2904
no
関連キーワード分類処理は
終了したか？
303
yes
細分類は
できるか？
304
no
yes
細分類処理B
2905
yes
未処理フォルダは
あるか？
306
no
終了

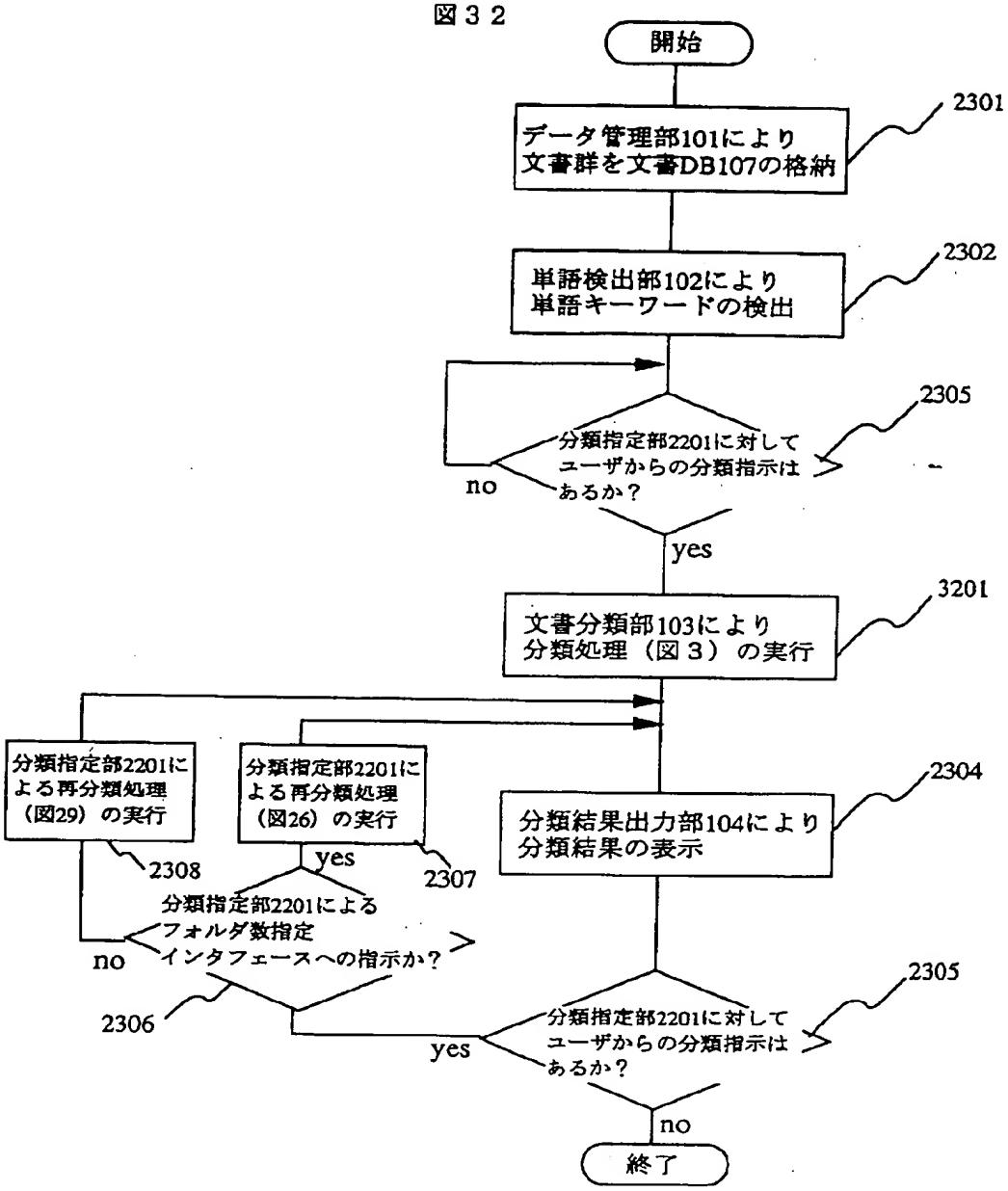
【図３１】

図３１


文書3100
3101
第一章
3102
キーワードaが
キーワードbに…
3105
3106
第二章
第三章
3103
第四章
3104
…キーワードgの…
キーワードh
3107
3108
108
キーワードDB
文書3100のキーワード群3110
キーワードa
キーワードb
キーワードc
キーワードe
キーワードf
キーワードg
キーワードh
3111
一時キーワードDB
作成
文書3100の一時キーワード群3112
キーワードa
キーワードb
キーワードg
キーワードh

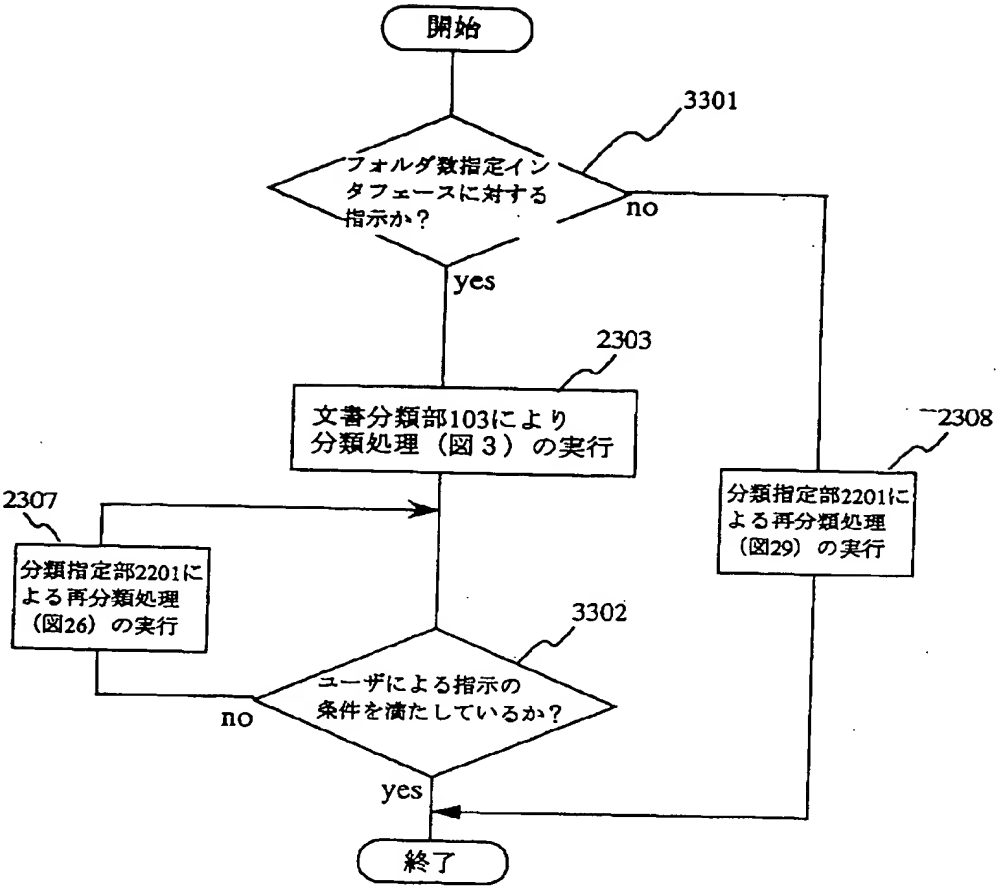
【図３２】

図３２


開始
データ管理部101により
文書群を文書DB107の格納
2301
単語検出部102により
単語キーワードの検出
2302
分類指定部2201に対して
ユーザからの分類指示は
あるか？
2305
no
yes
文書分類部103により
分類処理（図３）の実行
3201
分類指定部2201に
よる再分類処理
（図29）の実行
分類指定部2201に
よる再分類処理
（図26）の実行
分類結果出力部104により
分類結果の表示
2304
2308
yes
2307
分類指定部2201による
フォルダ数指定
インタフェースへの指示か？
no
2306
yes
分類指定部2201に対して
ユーザからの分類指示は
あるか？
2305
no
終了

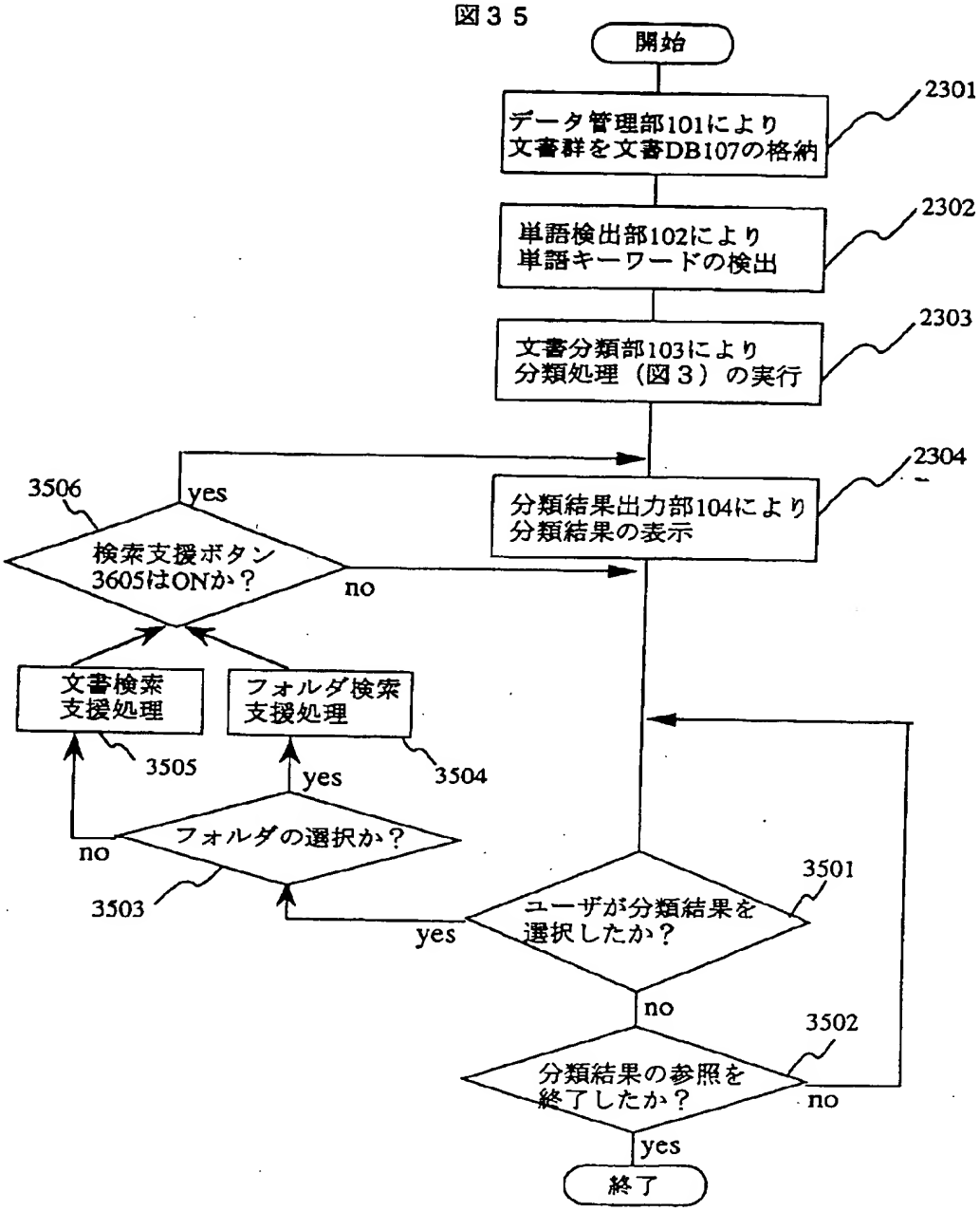
【図３３】

図３３


開始
3301
フォルダ数指定インタフェースに対する指示か？
no
yes
2303
文書分類部103により分類処理（図３）の実行
2308
分類指定部2201による再分類処理（図29）の実行
2307
分類指定部2201による再分類処理（図26）の実行
3302
ユーザによる指示の条件を満たしているか？
no
yes
終了

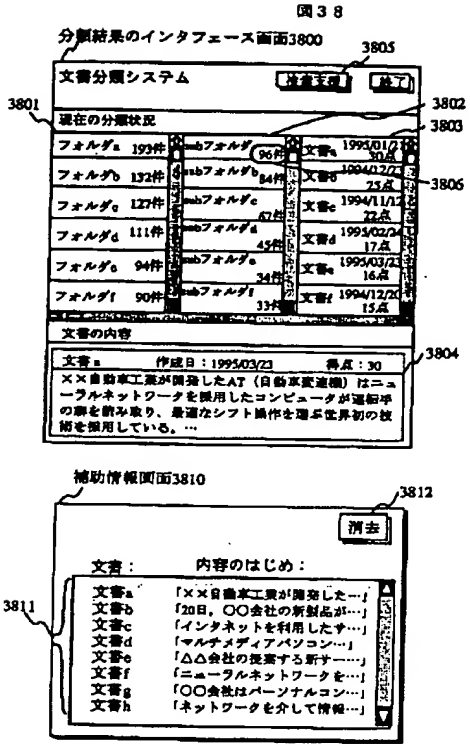
【図３５】

図３５


開始
データ管理部101により
文書群を文書DB107の格納
2301
単語検出部102により
単語キーワードの検出
2302
文書分類部103により
分類処理（図３）の実行
2303
分類結果出力部104により
分類結果の表示
2304
3506
yes
検索支援ボタン
3605はONか？
no
文書検索
支援処理
フォルダ検索
支援処理
3505
yes
3504
フォルダの選択か？
no
3503
3501
ユーザが分類結果を
選択したか？
yes
no
3502
分類結果の参照を
終了したか？
no
yes
終了

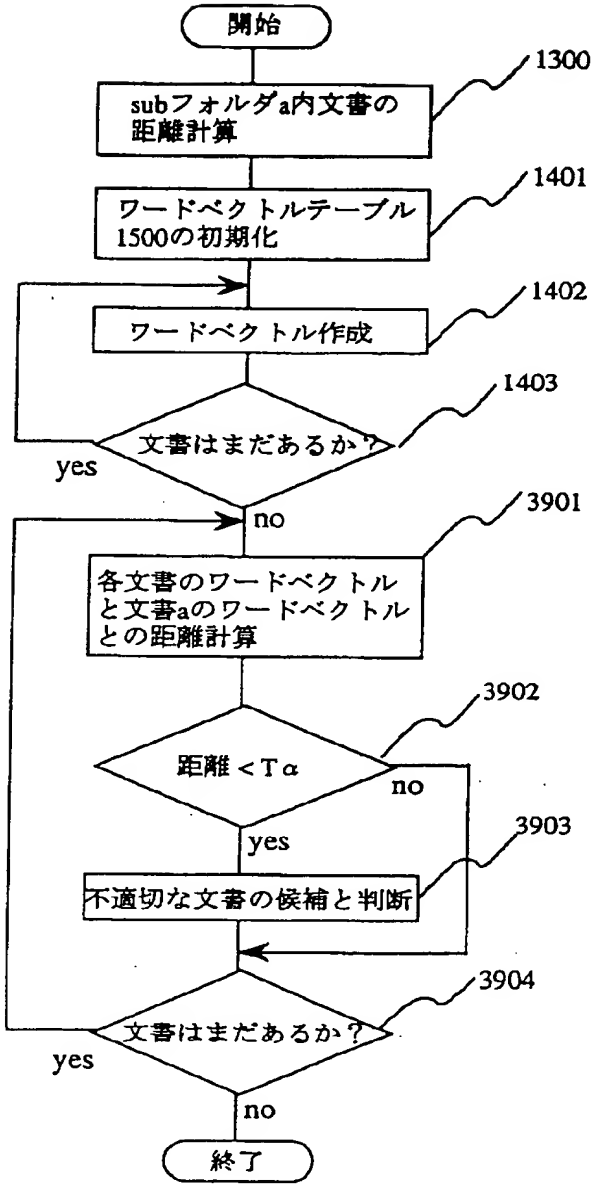
【図３８】

図３８



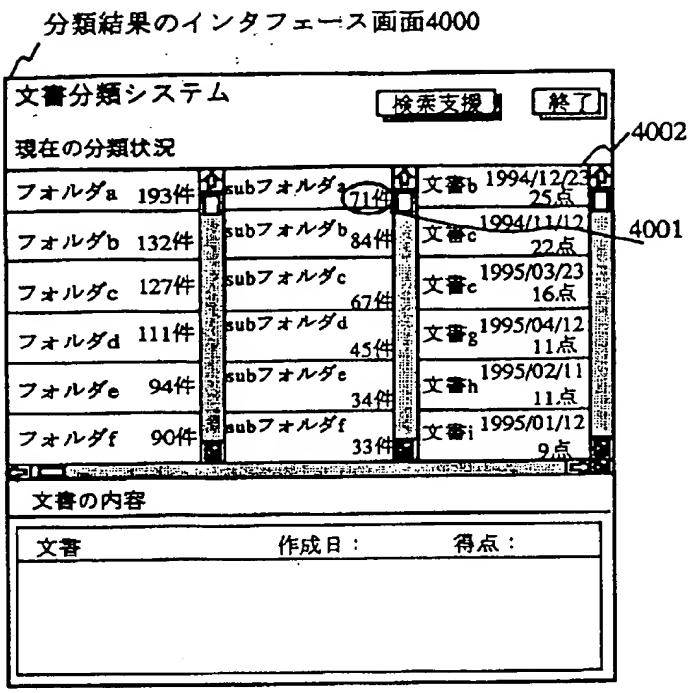
【図３９】

図３９



【図４０】

図４０



フロントページの続き

(72)発明者　橋本　哲也
神奈川県川崎市麻生区王禅寺1099番地　株
式会社日立製作所システム開発研究所内

An Automatic Document Classification Method Based on a Semantic Category Frequency Analysis

by Atsuo Kawai

Summary:

Conventional automatic document classification methods can be classified into two groups: (1) a method using statistical information of word notations or kanji notations, and (2) a method using language information and knowledge dependent on a classification system. Although method (1) has high generality due to the independence of its processing from its classification system, it has low precision because it uses only superficial information of word notations, etc. On the other hand, although method (2) has high precision because it uses information dependent on its classification system, it has low generality. This paper proposes a method of applying method (1) to a semantic category system generated independently of its classification system. In this method, first, a system automatically learns a semantic category which is biased and appears in each category field, based on a category example (a

document and its classification system). Then, unclassified documents are classified by using this learning result. In this method, generality can be secured in a sense that even when the classification method is changed, it is sufficient only if a corresponding category example (document for learning) is prepared, and if there is no need to modify the semantic category system used. When its recall ratio is compared with that of a conventional method using only word notations in an experiment where newspaper articles are targeted, the recall ratio has been improved by 10 to 12 %, thereby verifying its effectiveness.

1. Introduction

In this paper automatic document classification is studied. There are two document classification methods that can be used with a computer: a method of classifying documents into category fields prescribed in advance, and a method of automatically generating classification system frameworks from given unclassified document groups called a clustering. This paper studies the former method.

The method of classifying documents into category

fields prescribed in advance is further roughly divided into two groups: (1) a method using statistical processing (independent of a classification system), and (2) a method using language information and knowledge dependent on a classification system.

In method (1), a system automatically learns words, etc., which are biased and appear in each category field (hereinafter called field-distinguished words), based on the frequency statistics words and kanji notations in a document already classified. Then, the category of the document is determined based on the field-distinguished words in an unclassified document. Since, even when a classification system is changed, it is sufficient only if corresponding classified sample data have been prepared, this method has high generality. However, since only superficial information of word notations, etc., are used in this method, it has low precision.

In method (2), a classification system to be processed is fixed, and classification is performed by using information dependent on the classification system. To be more specific, it includes a method in which classification criteria are incorporated in a system by using generation rules, and a method of

classifying by following hierarchical correspondence tables between keywords, key concepts, and classification fields. Although this method has high precision, it has low generality, since the generation rules and correspondence tables between keywords and classification fields have to be entirely modified manually according to the extension and change of classification system.

Therefore, in this paper, a method utilizing a thesaurus prepared independently of a classification system is proposed in order to secure generality for a classification system and to improve classification precision by using non-superficial information.

In Chapter 2 automatic document classification is studied, and in Chapter 3 the results of experiments using this method are studied, including a comparison with conventional methods.

3. Experiments and Study

3.3 Evaluation criteria

A conformity ratio (indicating the size of a classification noise) and a recall ratio (indicating the size of classification omission) are used as

evaluation criteria, as in the case of information retrieval.

$$\text{Conformity ratio} = \frac{\text{RETrel}}{\text{RET}} \cdot 100$$

$$\text{Recall ratio} = \frac{\text{RETrel}}{\text{rel}} \cdot 100$$

where

RET = Total of category candidates outputted by a system (in all documents)

rel = Total of correct categories

= Total number of all documents to be processed (Note: This is because one document is provided with one correct category.)

RETrel = Total of correct categories outputted by a system

If a high threshold is set in equation (5) described in section 2.2.3, a classification field candidate with high reliability is focussed on, its conformity ratio increases, and its recall ratio decreases. Conversely, if a low threshold is set, the conformity ratio decreases, and the recall ratio increases. Therefore, in an experiment a threshold is not fixed and is variously set between 10 and 50, and the relation between the recall ratio and conformity ratio is examined.

3.4 Evaluation of precision

Figs. 5.1 and 5.2 show the evaluation of the relation between a conformity ratio and a recall ratio in classification systems 1 and 2, respectively. Compared with the conventional case, where only a word dictionary is used, the recall ratio and conformity ratio are better by 12(10)% and 9(10)%, respectively, when both a semantic category dictionary for classification and a word dictionary are used (figures outside paper theses and those inside parentheses are for systems 1 and 2, respectively, and threshold Th in equation (5) is assumed to be 25 in both systems).

When compared with the case where only a semantic category dictionary is used, higher precision is obtained when both a semantic category dictionary and a word dictionary are used. (The obtained values are at a maximum in systems 1 and 2 when t0=0.5 and t0=0.2, respectively.) The reasons are studied using "JR" and "KDD" for words and "transportation and traffic" and "information and communication" for classification fields. When a word dictionary is used, for example, two fields can be distinguished by using the word dictionary in the form of "JR (transportation and traffic)" and "KDD (information and

communication)". However, since in the semantic category system used in this experiment, both "JR" and "KDD" are only provided with a "company name" semantic category, the two fields cannot be distinguished. In order to distinguish between these two fields using a semantic category system, more detailed semantic categories have to be established. That is, in this example, the resolution of a semantic category system is too broad for the resolution of category fields. In such a case, a word dictionary is effective, and it improves precision.

In this case, the precision can also be improved by classifying a thesaurus (semantic category system) in more detail. When this customization is compared with the manual composition of a field-dependent thesaurus, with one knowledge expression dependent on the classification system described in method (2) in Chapter 1, there is no difference between them in individual jobs. An example of this is in the setting of the "communication" and "railway transportation" classification fields shown in Fig. 6 and the determination of company names belonging to classification fields.

However, the method studied in this paper is different from method (2) described in Chapter 1, in

the point that an objective score weighting to individual semantic categories due to learning using frequency information is available after the detailed classification of the semantic category system. For example and as shown in Fig. 6, it is assumed that the business field of corporations is classified as a semantic category. In this case, "air transportation" often appears in articles on the transportation and communication field, and it also appears in articles on the diplomacy field. In this way, a "classification field" corresponds to fields in which it appears, one to n, and the degree of its individual relation varies with the classification field to which it belongs. According to the method studied in this paper, this degree of relation (score) can be statistically calculated using a document for learning. On the other hand, in method (2) described in Chapter 1, only semantic categories and words that appear frequently are registered for each field. In order to take into consideration the degree of relation between a classification field and a semantic category or word, a manual intuitive degree of relation must be used.

In method (2) described in Chapter 1, when a classification system is partly modified after customization, corresponding modification is required.

However, in this method, since the score weighting to a semantic category is modified by new learning, there is no need to modify the thesaurus. As an example of this, the classification fields shown in the upper column of Fig. 6 are modified to those shown in the lower column of Fig. 6. In this case, the relation between a classification field newly set and the existing semantic category (including the semantic category shown in Fig. 6) has to be newly established. In method (2) described in Chapter 1, this setting has to be performed manually. However, in the method described in this paper, this relation can be automatically established by using a document for learning based on a new classification system.

There are also documents in which a correct category cannot be obtained, even if both a word dictionary and a semantic category dictionary are used. Out of these documents, three factors, from the viewpoint of word notation and semantic category, are focussed on in this paper, as given below.

(1) There is a document in which no correct category can be obtained because words contained in a document cannot be grouped at the semantic category level, although they can be grouped in a certain point of view. For example, as words which appear in a

document, there are "an unearthed article" ("article"), "an excavation" ("gathering") and "a cultural asset" ("asset"). These words can be grouped in an archaeology classification (field=culture). However, at the semantic category level the words can only be supplied with semantic categories, "article", "excavation" and "asset", and it is difficult to extract a relation, field=culture (archaeology).

(2) Errors result due to processing words with a polysemy (a plurality of semantic categories), while their semantic categories are not narrowed.

(3) Errors result due to adding the score of category candidates, taking into consideration the individual word notation and semantic category of words composing an idiom, etc. (for example, grass root). For example, grass-root personal computer communication: agriculture and forestry and fishery (incorrect category), information communication (correct category).

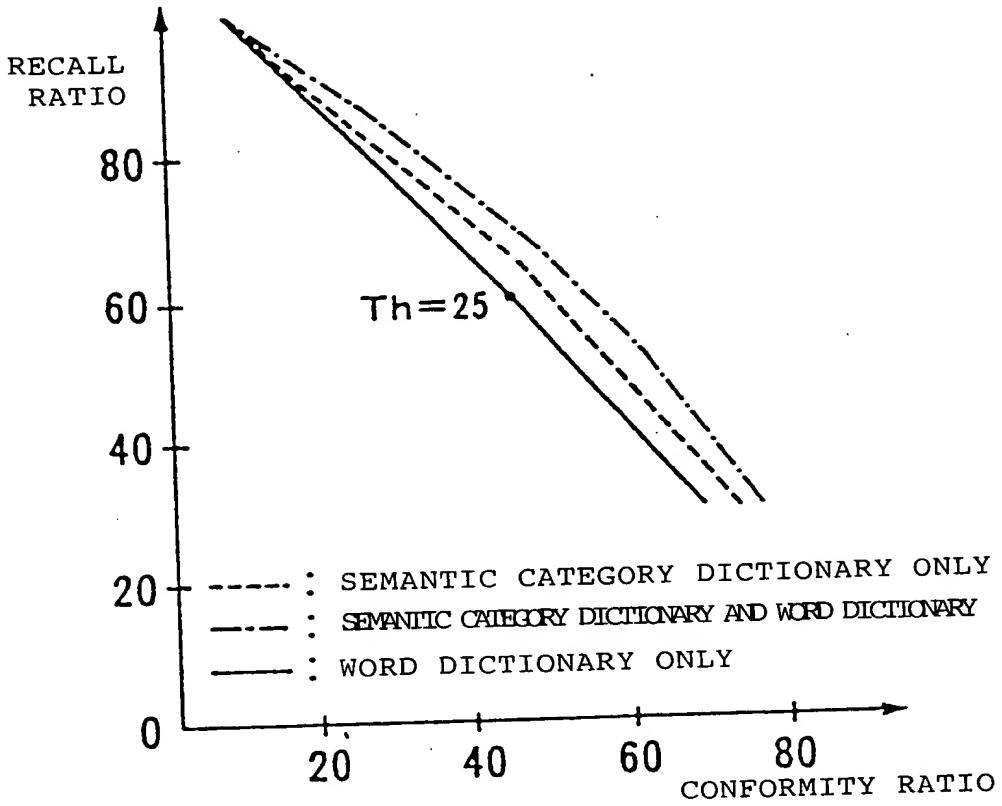


Fig. 5-1 Evaluation of reference dictionaries (classification system 1) (Recall Precision Graph).

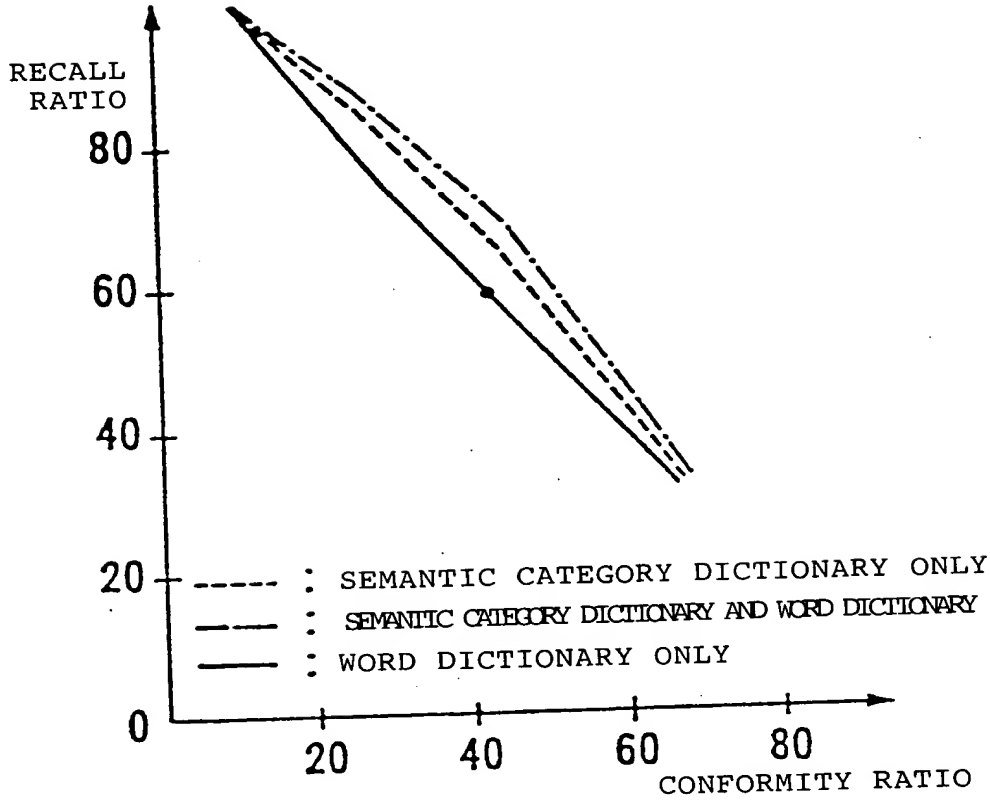


Fig. 5-2 Evaluation of reference dictionaries (classification system 2). (Recall Precision Graph).

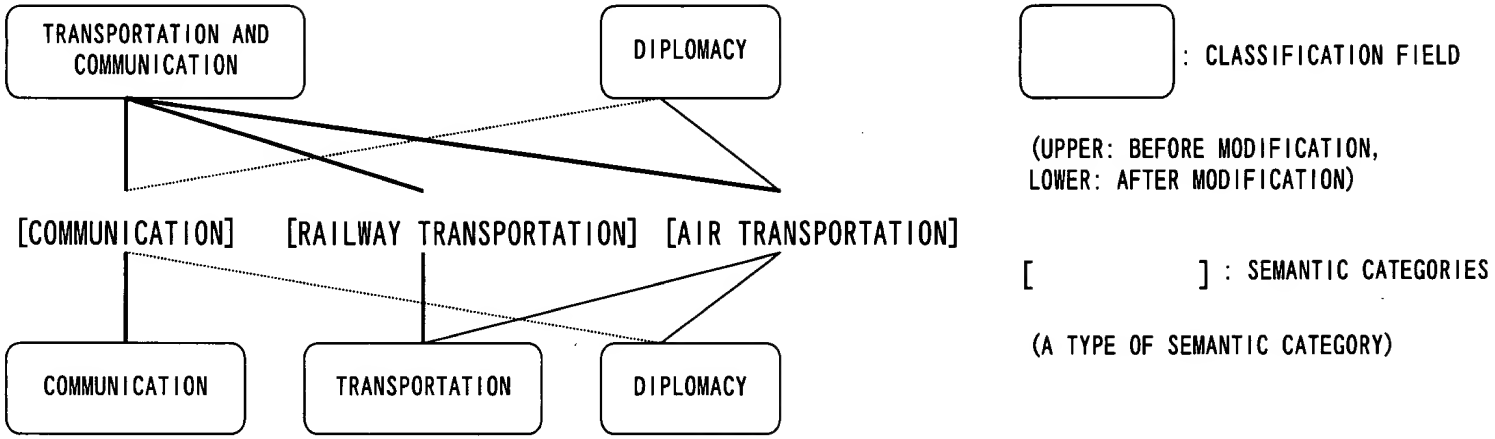


FIG. 6 Relation between classification fields and semantic categories.